(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

23

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 February 2002 (14.02.2002)

PCT

(10) International Publication Number
WO 02/13522 A2

(51) International Patent Classification[7]: H04N 5/76

(21) International Application Number: PCT/US01/41626

(22) International Filing Date: 7 August 2001 (07.08.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/637,311 10 August 2000 (10.08.2000) US

(71) Applicant: **QUINDI** [US/US]; 480 California Avenue, Suite 304, Palo Alto, CA 94306 (US).

(72) Inventors: **ROSENSCHEIN, Stanley, J.**; 907 Cowper Street, Palo Alto, CA 94301 (US). **GARAKANI, Arman**; #2, 31 Upland Road, Cambridge, MA 02140 (US). **BIRCHFIELD, Stanley, T.**; 1680 Los Padres Boulevard, Santa Clara, CA 95050 (US). **GILLMOR, Daniel, K.**; 3538 18th Street, Apt. 7, San Francisco, CA 94110-1636 (US).

(74) Agents: **VAN GIESON, Edward, A.** et al.; Fenwick & West LLP, Two Palo Alto Square, Palo Alto, CA 94306 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/13522 A2

(54) Title: AUDIO AND VIDEO NOTETAKER

(57) Abstract: A system and a method dynamically captures and analyzes audio, video, and event annotation data in a convenient and unobtrusive manner. An audio and video notetaker system captures audio, video, and additional event indicators, or "bookmarks". The audio and video notetaker system analyzes the captured data and correlates bookmarks with the audio and video data. The audio and video notetaker system provides real-time data analysis wherein the real-time analysis generates a model of events occurring in the recorded data. The audio and video notetaker system stores the audio, video, and analysis data in a format adapted for random access.

AUDIO AND VIDEO NOTETAKER

INVENTORS

Stanley J. Rosenschein, Arman Garakani, Stanley T. Birchfield, and Daniel Kahn Gillmor.

BACKGROUND

Field of Invention

The present invention relates generally to audio and video capture, and more particularly, to creating in real time a randomly accessible digital recording of audio and video including bookmarks and other metadata.

Background of the Invention

Much of the work of organizations is done in small-group meetings: executive or managerial meetings, design sessions, interviews, planning meetings, training sessions, focus groups, and so on. The aggregate investment made by corporations in such meetings is very large. Personal meetings and discussions still provide one of the most effective formats for generating and sharing information and making decisions.

The real value of a meeting, however, often lies not in what happens during the meeting itself, but in what happens afterwards. The ideas, information, and decisions generated during a meeting often support an entire network of follow-on activities: production of reports, design specifications, presentations, and action items, as well as communications with team members, supervisors, colleagues, and clients. Ultimately, an organization's return on investment in meetings is measured by how well meeting outputs support follow-on execution.

The success of these follow-on efforts is largely dependent on the documentation of the content of the meeting. Current methods for recording information during a meeting and disseminating that information have numerous disadvantages. Human memory is fallible and it is often impossible to consistently verify or confirm the contents of a meeting from individual memories after a meeting has ended. Handwritten or typed notes are brief and often

indecipherable. Furthermore, the note-taking process itself tends to distract a meeting participant and prevent him or her from fully participating in an ongoing meeting.

Existing systems for recording the audio and visual aspects of meetings also have disadvantages. A taped audio recording is time-consuming to replay. Taped audio is not randomly accessible, and it is generally necessary to listen to a large percentage of a meeting a second time in order to review decisions made during the meeting or find other pertinent information. Further, the tape records only what was actually said during the meeting; it does not allow individual meeting participants to annotate particular items of interest without verbally expressing them. Audio tape fails to record important information such as speakers' facial expressions, non-verbal actions and other cues from meeting participants. It may also be difficult to distinguish who is speaking at any given time on the meeting audio tape.

Video recordings of meetings similarly require a great deal of time to review, and to easily and reliably capture. Current videotaping systems merely allow for an entire meeting to be replayed, without providing a system for easily and quickly skipping between speakers and events. Meeting participants are also unable to annotate events onto the videotape without interrupting the meeting with visual or verbal cues. Furthermore, the effect of a large video camera recording a meeting can be burdensome and uncomfortable, and may itself cause participants to suppress their comments or contributions. Videotaping often requires extra lighting, cumbersome and distracting cameras, and a camera operator to ensure reasonable video quality.

A system is needed for providing an easily reviewable video and audio recording of a meeting or other interpersonal communication. The system should be easy to operate and unobtrusive during the meeting. The system should provide a way for meeting participants to annotate the recording in regard to events during the meeting and easily refer back to those events during meeting review. The system should store the audio, video, and event annotation data in a manner facilitating replay immediately after recording and allowing the stored meeting to be easily and quickly disseminated to others.

SUMMARY OF THE INVENTION

The present invention allows for the dynamic capture and analysis of audio, video, and event annotation data in a convenient and unobtrusive manner. An audio and video notetaker system receives audio, video, and additional event indicators, or "bookmark" signals, for example, during a meeting. The audio and video notetaker system analyzes the captured data

and associates bookmark data with the audio and video data. The audio and video notetaker system provides real-time data analysis to generate a model of events occurring in the recorded data. The audio and video notetaker system stores the audio, video, and analysis data in a format adapted for random access.

In one embodiment, the audio and video notetaker is a portable device dimensioned to be easily carried within a briefcase or similar bag. The audio and video notetaker comprises a plurality of microphones and one or more cameras providing multidirectional video recording. The audio and video notetaker includes one or more ports for receiving event indication signals. Event signals are generated by a portable bookmarker that may be communicatively coupled to the notetaker via the ports.

In another embodiment, the audio and video notetaker system is connected to a port extender that includes additional ports for receiving additional data input.

The features and advantages described in the specification are not all-inclusive, and particularly, many additional features and advantages will be apparent to one of ordinary skill in the art in view of the drawings, specification, and claims hereof. Moreover, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the inventive subject matter, resort to the claims being necessary to determine such inventive subject matter.

The foregoing merely summarizes aspects of the invention. The present invention is more completely described with respect to the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1A is an illustration of a meeting held using an embodiment of an audio and video notetaker system.

Fig. 1B is an illustration of a meeting held using another embodiment of an audio and video notetaker system.

Fig. 1C is an illustration of a meeting held using another embodiment of an audio and video notetaker system.

Fig. 1D is an illustration of a timeline for a meeting using an audio and video notetaker system.

Fig. 2A is an illustration of the overall design of an embodiment of an audio and video notetaker.

Fig. 2B is an illustration of the external features of an embodiment of an audio and video notetaker.

Fig. 3 is a block diagram illustrating the internal components of an embodiment of an audio and video notetaker.

Fig. 4 is a block diagram illustrating the internal components of an embodiment of a port extender.

Fig. 5 is a block diagram illustrating the internal components of an embodiment of a manual bookmarker.

Fig. 6 is a block diagram illustrating an embodiment of an audio and video notetaker PC version mini docking station.

Fig. 7 is a flowchart illustrating an embodiment of a method for processing and storing data from an audio and video notetaker system.

Fig. 8 is a block diagram illustrating an embodiment of a method for selective video capture in an audio and video notetaker system.

Fig. 9 is a block diagram illustrating an embodiment of the functionality of an analyzer in an audio and video notetaker system.

The figures depict a preferred embodiment of the present invention for purposes of illustration only. One skilled in the art will readily recognize from the following discussion that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the invention described herein.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference will now be made in detail to several embodiments of the present invention, examples of which are illustrated in the accompanying drawings. Wherever practicable, the same reference numbers will be used throughout the drawings to refer to the same or like parts. The figures depict preferred embodiments of the present invention for purposes of illustration only. One skilled in the art will readily recognize from the following discussion that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the invention described herein.

Fig. 1A is an illustration of a meeting held using an embodiment of an audio and video notetaker system. An audio and video notetaker 100 is placed in a centralized location during the meeting and creates an audio and video record of the meeting. The meeting is conducted under ordinary conditions, without special lighting or staging requirements. Typical ambient lighting 130 is provided, and the audio and video notetaker 100 is placed on an ordinary table 140.

Meeting participants A, B, C and D conduct the meeting and move about the room without needing to alter their actions due to the presence of the audio and video notetaker 100. For example, participant D moves about the room to point to a bulletin board 124, while participants A-C remain seated around the table 140. The meeting participants A-D may freely move in a preferably 360° circumference range surrounding the audio and video notetaker 100 during the meeting, and they will continue to be captured in the audio and video record of the meeting.

The audio and video notetaker 100 records audio and video during the meeting. The audio recording is multidirectional, and is capable of capturing audio from all of the participants in the meeting. The video recording is also multidirectional, and similarly is capable of capturing video of all of the participants, in one example of approximately a 360° view. The video recording device within the audio and video notetaker 100 is positioned to allow the video capture of all participants, including, for example, facial motions, chart 124, and other presentation items displayed during the meeting. The video recording is not blocked by the presence of typical meeting objects, such as coffee cups 126, provided they are sufficiently set back from the audio and video notetaker 100.

In the embodiment shown in Fig. 1A, the audio and video notetaker system includes an audio and video notetaker 100 and bookmarking devices 112A and 112B. The audio and video notetaker 100 operates in a "stand-alone" mode and is conveniently sized to be portable. As will be described in more detail in the following description, the audio and video notetaker 100 contains all of the components required to perform audio and video recording of a meeting.

The bookmarking devices 112 are hand-held devices by which meeting participants transmit signals to the audio and video notetaker 100. Signals from the bookmarking devices 112 are received and interpreted by the audio and video notetaker 100 as indicating an event of interest has occurred during the meeting. In one embodiment, signals from the bookmarking devices may also direct the audio and video notetaker 100 to pause or resume the meeting recording.

As shown in Fig. 1A, the audio and video notetaker 100 is a small, portable device. The audio and video notetaker 100 is sized to fit conveniently in a briefcase or other hand-held bag, and typically is smaller than a pie tin. The audio and video recording devices within the audio and video notetaker 100 are unobtrusive, and do not distract participants from the meeting.

The audio and video notetaker 100 performs real-time digital audio and video recording of a meeting. Bookmarks indicating events of interest may also be added to the audio and video notetaker 100 meeting record through the bookmarkers 112. For example, when a particular decision is made, a significant comment or the like, a user may click a button on the bookmarker 112 to indicate the event. The audio and video notetaker 100 also performs real-time analysis and processing on the recorded video and audio data, including incorporating bookmark data into the recorded meeting data. Upon completion of the meeting, a stored digital record of the meeting exists in the audio and video notetaker 100. This stored meeting record may be easily edited and annotated by the audio and video notetaker owner on a personal computer, using audio and video notetaker editing software. Additionally, copies of the edited meeting record may be sent to any number of other interested parties, such as meeting participants, absentees from the meeting, and people wishing to review the meeting. The edited meeting record is easily viewed by others using audio and video notetaker viewing software.

Fig. 1B is an illustration of a meeting held using of another embodiment of an audio and video notetaker system. In this embodiment, the audio and video notetaker 100 is connected to a port extender 110. The port extender 110 includes additional ports for receiving additional external inputs. For example, the port extender 110 allows the audio and video notetaker 100 to receive computerized slide presentations, electronic whiteboard signals, teleconferencing audio and video, and pre-recorded audio and video feeds. The port extender 110 operates in a pass through manner, i.e., signals are input into the audio and video notetaker 100 for storage and processing, and may still be sent to other devices used in the meeting itself. This allows the audio and video notetaker 100 to capture and record an electronic slide show while the slide presentation is being shown in a meeting.

Fig. 1C is an illustration of a meeting held using another embodiment of an audio and video notetaker system. In this embodiment, the audio and video notetaker 100 is connected to a computer 128. The computer 128 may be a laptop-style computer or a desktop model. Meeting participant A transmits signals to the audio and video notetaker 100 using a bookmarking device 112A. Meeting participant B also transmits signals to the audio and video notetaker 100 using the computer 128. The computer 128 may be used to provide storage space for audio and video recording, or as additional storage space supplementing the storage within the audio and video notetaker 100 itself.

The computer 128 may also be used as a notetaking utility accompanying the meeting. Participant B types in notes and comments that are stored along with the audio and video meeting record. Participant B additionally may create customized bookmark signals corresponding to different computer keystroke combinations, mouse clicks, or other signals triggered through the utility's graphical user interface.

Fig. 1D is an illustration of a sample timeline for a meeting using an audio and video notetaker system. The timeline indicates the timing of actions taken by each of five different participants A, B, C, D and E. Participants A-D participate to different degrees in the meeting itself. Participant E does not participate in the meeting, but participant E later receives an annotated copy of the meeting recording.

The steps shown in Fig. 1D apply to the use of either an audio and video notetaker 100 with or without the port extender 110, referred to collectively herein as an audio and video notetaker system. The audio and video notetaker 100 may additionally be connected to a personal computer or a local area network (LAN). Such a connection may be wired (for

example, an Ethernet connection) or a wireless configuration (such as a wireless network card).

In the meeting of Fig. 1D, both participant A and participant B have bookmarking devices 112. These bookmarking devices 112 may be dedicated bookmarkers for use with the audio and video notetaker system. In another embodiment, participant A may use a computer 128 connected to the audio and video notetaker 100 to send bookmarking signals to the audio and video notetaker 100. In still another embodiment, participants A or B may use a hand-held computing device as a bookmarker 112, where the hand-held computing device is adapted to send signals to the audio and video notetaker system. In the following, it is assumed that a bookmarker 112 is used.

At time t_0, participant A attaches the audio and video notetaker 100 to a computer 128 if a computer attachment is desired. In another embodiment, the audio and video notetaker 100 is used in a stand-alone mode, or a port extender 110 is used either with or without the computer 128. At time t_1, participant A turns on the audio and video notetaker system, and the audio and video notetaker system begins recording audio and video. At time t_2, participants A, B and C begin the meeting. The meeting progresses normally, and the presence of the audio and video notetaker system device does not necessitate any modifications to the meeting room or to the behavior of the meeting participants.

At time t_3, participant A bookmarks an event of interest. Participant A uses the bookmarker 112 to send a bookmark signal to the audio and video notetaker system. In response to the bookmark signal, the audio and video notetaker system saves data representing the bookmark into the real-time digital meeting recording. The data corresponding to the bookmark signal will be saved in the meeting recording, indexed to the recording time at which the signal was sent, in order to flag the event of interest during meeting playback. In one embodiment, bookmark signals sent from different bookmarkers 112 are identical, and are treated in the same manner by the audio and video notetaker system. In another embodiment, different bookmarkers 112 each have a signal value, allowing the audio and video notetaker system to distinguish between a bookmark sent from participant A and a bookmark sent, for example, from participant B.

At time t_4, participant A moves to the whiteboard in the meeting room. The audio and video notetaker system is able to continue to capture audio and video of participant A, because the audio and video notetaker system records both audio and video of the entire meeting room.

In one embodiment (where the audio and video notetaker system has whiteboard input), the audio and video notetaker system is also able to capture writing on the whiteboard in two separate ways. The audio and video notetaker system records video over a 360° field of view, and thus the video recording will include the written whiteboard text or diagrams, though at coarse resolution. However, an electronic whiteboard may also be plugged into an audio and video notetaker 100 via a port extender 110. In this manner, the electronic whiteboard may download captured information directly into the audio and video notetaker system.

At time t_5, participant B bookmarks an event. At time t_6, a new meeting participant, participant D, enters the meeting room. At time t_7, participant A takes the meeting "off the record" thereby pausing the audio and video recording. Participant A may use a bookmarking device 112 to take the meeting off the record, or may tap the audio and video notetaker system on/off ring itself to perform the desired action. Taking a meeting "off the record" may be used, for example, to discuss confidential or sensitive information. At time t_8, participant A resumes recording by putting the meeting back "on the record."

At time t_9, participants A, B, C and D end the meeting. At time t_{10} participant A turns off the audio and video notetaker system and unplugs the audio and video notetaker 100 from the personal computer 128 if a personal computer 128 is in use. A stored digital recording of the meeting now exists in the audio and video notetaker system. It is indexed with all of the saved bookmark data, indexed to the time of the bookmarked event and preferably including identifying indicia. Processing and analysis were performed on the audio, video and bookmarks in real time while the meeting was in progress, and the analysis information is stored as metadata. The stored digital recording may be easily edited and distributed to others.

At time t_{11}, participant A edits the stored meeting recording, adding text notes and annotations to selected portions of a recording, to flag items of interest, comment on various events, decisions, or statements, etc. The stored meeting recording is transferred from the audio and video notetaker system to a personal computer using an audio and video notetaker mini docking station 600 or other type of audio and video notetaker system connection. In another embodiment, the stored meeting recording is transferred during the meeting to a computer 128 or to an external server system via the port extender 110, or via an attached personal computer, which is also connected to a LAN or other network system. Participant A edits the stored meeting recording on a personal computer using audio and video notetaker system editing software. Editing of the video is optional, as the stored audio and video data is usable in its raw format as well.

In one embodiment (where the audio and video notetaker system has whiteboard input), the audio and video notetaker system is also able to capture writing on the whiteboard in two separate ways. The audio and video notetaker system records video over a 360° field of view, and thus the video recording will include the written whiteboard text or diagrams, though at coarse resolution. However, an electronic whiteboard may also be plugged into an audio and video notetaker 100 via a port extender 110. In this manner, the electronic whiteboard may download captured information directly into the audio and video notetaker system.

At time t_5, participant B bookmarks an event. At time t_6, a new meeting participant, participant D, enters the meeting room. At time t_7, participant A takes the meeting "off the record" thereby pausing the audio and video recording. Participant A may use a bookmarking device 112 to take the meeting off the record, or may tap the audio and video notetaker system on/off ring itself to perform the desired action. Taking a meeting "off the record" may be used, for example, to discuss confidential or sensitive information. At time t_8, participant A resumes recording by putting the meeting back "on the record."

At time t_9, participants A, B, C and D end the meeting. At time t_{10} participant A turns off the audio and video notetaker system and unplugs the audio and video notetaker 100 from the personal computer 128 if a personal computer 128 is in use. A stored digital recording of the meeting now exists in the audio and video notetaker system. It is indexed with all of the saved bookmark data, indexed to the time of the bookmarked event and preferably including identifying indicia. Processing and analysis were performed on the audio, video and bookmarks in real time while the meeting was in progress, and the analysis information is stored as metadata. The stored digital recording may be easily edited and distributed to others.

At time t_{11}, participant A edits the stored meeting recording, adding text notes and annotations to selected portions of a recording, to flag items of interest, comment on various events, decisions, or statements, etc. The stored meeting recording is transferred from the audio and video notetaker system to a personal computer using an audio and video notetaker mini docking station 600 or other type of audio and video notetaker system connection. In another embodiment, the stored meeting recording is transferred during the meeting to a computer 128 or to an external server system via the port extender 110, or via an attached personal computer, which is also connected to a LAN or other network system. Participant A edits the stored meeting recording on a personal computer using audio and video notetaker system editing software. Editing of the video is optional, as the stored audio and video data is usable in its raw format as well.

making the camera appear less intrusive. Additionally, in one embodiment, the paraboloidal mirror is encased in a substantially cylindrically shaped portion of the audio and video notetaker 100. This substantially cylindrically shaped portion may be raised above the body of the audio and video notetaker 100 during video recording, and lowered back into the body of the audio and video notetaker 100 after recording is completed.

In another embodiment of multidirectional camera system 210, a 360° circumferential view camera system is implemented using a set of individual cameras and an upside-down pyramid-shaped mirror. The base of the pyramid-shaped mirror is located above the cameras, while the apex of the pyramid is located between the cameras. The sides of the mirrors each reflect different views from different directions downward onto separate cameras, which combine to provide a substantially 360° field of view. Each mirrored pyramid side has a separate camera, for example, a rectangle-based pyramid would have four separate cameras. This type of pyramid-shaped mirrored camera arrangement is described in both U.S. Patent No. 5,745,305, entitled "Panoramic Viewing Apparatus," issued April 28, 1995, and U.S. Patent No. 5,793,527, entitled "High Resolution Viewing System," issued August 11, 1998, both of which are incorporated by reference herein in their entirety.

In yet another embodiment of multidirectional camera system 210, four individual cameras are placed in a substantially ring-shaped arrangement around the body of the audio and video notetaker 100. Each individual camera is embedded into the body of the audio and video notetaker 100 and faces outward through a porthole (not shown) in the side of the audio and video notetaker 100. It will be evident to one of skill in the art that more or fewer individual cameras may be used. Each individual camera is positioned to view and record a portion of a substantially 360° field of view. The 360° field of view will be divided substantially equally between each of the individual cameras, with a number of cameras as needed given the angled of each camera's field of view.

The cameras comprising multidirectional camera system 210 may be arranged to provide individual views that partially overlap each other. Additionally, a camera may be added that is not part of a 360° field of view. For example, a separate camera may be set to record an individual speaker remotely, or a camera may be included to record an individual meeting speaker in a separate field of view.

In one embodiment, one or more individual cameras of the camera system 210 are angled slightly upward from horizontal in relation to the base of the audio and video notetaker

100, in order to facilitate capturing the viewing region of interest. For example, assume the audio and video notetaker 100 is sitting on a horizontal tabletop. Each individual camera is positioned to place the camera centerline at an angle approximately 30° upward from the horizontal plane of the tabletop. Each individual camera is therefore capable of capturing images located in a range of approximately -8° to 68° upward from the horizontal plane of the tabletop. It will be evident to one of skill in the art that the angle of view for each individual camera may be varied as desired for different ranges of view. In another embodiment, the individual cameras are arranged to capture a view 360° in circumference and hemispherically shaped.

In yet another embodiment, a multidirectional camera system is set up to capture an "interview-style" field of view, for example, where two participants are seated across from each other. In this embodiment, two or more cameras are pointed in opposite directions, and thus are pointing in directions approximately 180° apart from each other. In yet another embodiment, a single wide-angle camera may be used to cover the field of view of interest during a meeting or other event. Additionally, multiple audio and video notetaker systems may be linked together to cover large or unusually shaped rooms. As used herein, the term "multidirectional" camera is understood to encompass camera systems covering a field of view suitable for recording a meeting or other type of event.

Four microphones 220A, 220B, 220C and 220D are distributed substantially equidistant from each other around the body of the audio and video notetaker 100. In one embodiment, the four microphones 220 are placed 90° from each other in a horizontal ring around the audio and video notetaker 100. The four microphones 220 provide multidirectional, synchronized four-channel audio. It will be evident to one of skill in the art that more or fewer microphones may be used. For example, in another embodiment, three microphones are placed 120° from each other in a horizontal ring around the audio and video notetaker 100. The number of microphones 220 must be sufficient to provide multi-channel audio suitable for use in triangulation software, for determining the originating physical location of a particular sound.

Ports 230 are also located on the outside of audio and video notetaker 100. The ports 230 are adapted to receive signals from bookmarking devices indicating that an event of interest has occurred in a meeting, or transmitting a command to the audio and video notetaker 100. The bookmarking devices may transmit ultrasonic, infrared, or radio signals in a wireless mode to the audio and video notetaker 100, in which case ports 230 include appropriate signal

detectors/receivers. In another embodiment, the bookmarking devices are attached to the audio and video notetaker 100 by a wire capable of transmitting the signals, in which case ports 230 provide suitable jacks. The ports 230 are adapted to receive the type of signal generated by the associated bookmarking device.

An LED 242 displays an indication of the current state of the audio and video notetaker 100. For example, if the audio and video notetaker 100 is "on," "paused," or "stopped," a light indicating one of these states will display on the LED 242. In another embodiment, a different type of display may be used, such as a LCD 240 for displaying messages. The audio and video notetaker 100 may contain different combinations of external display mechanisms, as will be evident to one of skill in the art.

An on/off tap ring 250 is a substantially ring-shaped portion of the audio and video notetaker 100 that when touched or tapped changes the state of the audio and video notetaker 100. The ring shape allows the tap ring 250 to be easily touched and activated from any direction. In one embodiment, the tap ring 250 is a visual indicator that glows with light when the audio and video notetaker 100 is in the "on" state and is recording audio and video, thereby providing a visual indication of the audio and video notetaker 100's state of operation. In another embodiment, the tap ring 250 is implemented as a button, which changes the audio and video notetaker 100's state when pressed.

Fig. 3 is a block diagram illustrating the internal components of an embodiment of an audio and video notetaker 100. All of the internal components are sized to fit within the portable audio and video notetaker 100 capable of being placed conveniently into a briefcase or other hand-held carrying case. As will be evident to one of skill in the art, the actual internal layout of the components may be arranged in any convenient manner.

The audio and video notetaker 100 receives and transmits a variety of signals via a camera input 310, a microphone audio input 320, and various other input and output ports. Data input and output is connected to a central processing unit (CPU) 340, a memory 350, and the removable storage media 360 via a system bus 380. Data processing is performed by the CPU 340 using the memory 350 within the audio and video notetaker 100. The resulting output is either stored within the audio and video notetaker 100 on the removable storage media 360, (or in another embodiment, a hard drive), or sent to an attached host device for additional processing and storage. Power is provided to the audio and video notetaker 100 from a battery, an A/C adaptor 390, or from attached hosts through ports.

The camera input 310 is connected to a video buffer controller 312. The video buffer controller 312 is connected to a video double-buffer 314. The camera input data 310 is temporarily stored in the double-buffer 314 and processing is performed on the video data. The video controller 312 transmits data and receives control instructions off of the system bus 380.

The microphone input 320 is connected to a group of analog-to-digital converters (ADC) 322, which converts the analog audio input signals 320 into digital signals. The converted audio input is received by an audio buffer controller 324 connected to an audio buffer 326. The audio data input 320 is temporarily stored in the buffer 326 and processing is performed on the audio data. The audio controller 324 transmits data and receives control instructions off of the system bus 380.

After processing, the audio data and video data are passed to a compression chip 370 for compression. In one embodiment, the compression chip is an MPEG2 compression chip. In one embodiment, the compressed audio and video MPEG2 data file is then transmitted to the system bus 380 for storage on the removable storage media 360. Alternatively, the compressed audio and video data is transmitted to a host universal serial bus (USB) port 334, where it is sent to a host such as an attached computer 128. In another embodiment, the host USB port 334 connects to an external hard drive or other storage mechanism for storing the data from the audio and video notetaker 100. Because the volume of data provided by the cameras and microphones in many cases will exceed the input resolution of the compression subsystem, the invention allows for "selective capture" of data, eliminating some data. Selective capture encompasses either uniform data subsampling or the selection and retention of "regions of interest" or any combination of the two.

A bookmark signal port 330 for receiving bookmark signals is adapted to receive one of either radio, ultrasonic or infrared bookmark signals. The bookmark signals port 330 is connected to the system bus 380. Bookmark signals are received and used to create associated metadata indicating the occurrence of a bookmark signal. The bookmark metadata are stored either on the removable storage media 360, or stored on a host device connected to the system bus 380 via the host USB port 334. The bookmark metadata may be stored as part of the audio and video compressed file, or may be stored as a separate file.

An auxiliary USB port 332 is adapted to receive additional audio and video notetaker 100 data. In one embodiment, the USB port 332 receives whiteboard signals. The auxiliary

USB port 332 is connected to the system bus 380. USB signals are received and stored either on the removable storage medium 360, or stored on a host device connected to the system bus 380 via the host USB port 334.

A high-performance serial bus port 336 is adapted to send and receive additional audio and video notetaker 100 data at data transfer speeds ranging from approximately 200-400 Mbps. The high-performance port 336 has sufficient bandwidth to carry multimedia audio and video signals, and is connected to both the video buffer controller 312 and the audio buffer controller 324 for providing additional video and audio data inputs. The high-performance port 336 is also connected to the system bus 380. In one embodiment, the high-performance port 336 is an IEEE 1394 serial bus port.

Fig. 4 is a block diagram illustrating the internal components of an embodiment of a port extender 110. The port extender 110 attaches to the audio and video notetaker 100 and extends the capabilities of the audio and video notetaker 100 by providing additional inputs and outputs to the audio and video notetaker 100. The port extender 110 includes a high-performance serial bus port 450 adapted to connect to the audio and video notetaker 100 high-performance serial bus port 336. In one embodiment, the high-performance port 450 is an IEEE 1394 serial bus port.

The port extender 110 receives additional audio and video signals in a pass through manner, enabling the audio and video notetaker 100 to receive a signal that is also being used or displayed during a meeting. For example, if an overhead projector is being used in a meeting, a computer generates the original video graphics array or super video graphics array signal (collectively referred to as "VGA" herein) of the presentation. The VGA computer signal is plugged into a VGA input port 410 on the port extender 110. The LCD display apparatus is plugged into a VGA output port 412, which is internally connected to the VGA input port 410 and allows the VGA signal to pass through the port extender 110. However, the VGA input port 410 is also connected to the high-performance serial bus port 450, which allows a host (the audio and video notetaker 100) connected to port 450 to also receive the VGA signal for processing and storage.

The VGA data is treated as an additional video input into the audio and video notetaker 100. However, the processing and storage of the VGA data will depend on the type of data being carried. High-bandwidth video will be stored as part of a compressed file along with the audio and video notetaker 100 camera inputs. However, if the VGA data is a series of lower

bandwidth static images, such as a slide show, the data will be stored as a set of discrete JPEG files or other type of image format, stored as metadata separately from the compressed files. A slide transition detector within the audio and video notetaker 100 detects an image change and triggers the storage of a new JPEG slide file.

The port extender 110 also includes a National Television Standards Committee (NTSC) input port 420 and an NTSC output port 422 for carrying a television or videocassette recorder signal. NTSC output port 422 is connected to NTSC input port 420 to allow signal pass through operations. NTSC input port 420 is also connected to the high-performance serial bus port 450. In another embodiment, ports 420 and 422 may be adapted to receive and transmit Phase Alternation Line (PAL) signals or Sequential Couleur avec Memoire (SECAM) signals. The NTSC signals are treated as an additional camera input into the audio and video notetaker 100 and stored as part of the audio and video compressed file.

The port extender 110 further includes two USB ports 430 and 432 suitable for carrying signals such as an electronic whiteboard signal. USB ports 430 and 432 are also connected to the high-performance serial bus port 450. Whiteboard signals are typically a discrete set of vectors, and will be stored as part of a separate metadata file.

The port extender 110 further includes an audio input port 440 and an audio output port 442 suitable for receiving audio signals, such as a teleconference phone line or an additional microphone input. Audio output port 442 is connected to audio input port 440 to allow signal pass through operations. Audio input port 440 is also connected to the high-performance serial bus port 450. The additional audio signals will be treated by the audio and video notetaker 100 as another microphone input, and are stored as part of the audio and video compressed data file.

As will be evident to one of skill in the art, numerous other types of input and output signals may be connected with ports on the port extender 110. The port extender 110 is capable of transmitting one or multiple different types of signals simultaneously to the audio and video notetaker 100 via the high-performance serial bus port 450. The number of signals that may be transmitted to the audio and video notetaker 100 simultaneously is limited only by the bandwidth of the high-performance serial bus. Relatively low-bandwidth application signals, such as whiteboard signals or slide images, may be transmitted in combination. Relatively higher bandwidth applications, such as a digital video signal, may require the full bandwidth of the high-performance serial bus.

Fig. 5 is a block diagram of an embodiment of the components of a manual bookmarking device 112. The bookmarker 112 contains two buttons for sending signals to an audio and video notetaker 100: a bookmark button 510, and an "off the record" button 512. The bookmark button 510 is pressed by a user to send a signal to the audio and video notetaker 100 indicating that an event of interest has occurred. The "off the record" button 512 is pressed by a user to send a signal to the audio and video notetaker 100 indicating that the audio and video notetaker operation should be paused. The audio and video notetaker 100 enters a non-recording state in which both audio and video recording is suspended. Button 512 may be pressed again to send a "restart" signal to the audio and video notetaker 100, returning the audio and video notetaker 100 to recording audio and video.

In one embodiment, bookmarker 112 also includes a light emitting diode (LED) 520 for providing feedback to a user. In one embodiment, the LED 520 flashes whenever button 510 or 512 is pressed. In another embodiment, LED 520 displays a lighted signal when the audio and video notetaker 100 is recording audio and video. For example, the LED 520 displays a green light to indicate that the audio and video notetaker 100 is turned on and recording, and displays a red signal to indicate that the device is paused and "off the record".

A transmitter 530 transmits signals from the bookmarker 112 to the audio and video notetaker 100. A receiver 532 receives signals back from the audio and video notetaker 100. Both the transmitter 530 and the receiver 532 may send and receive one of either ultrasonic, infrared or radio signals corresponding to the type of signals required by the audio and video notetaker 100. For example, in one embodiment, the transmitter 530 and receiver 532 send and receive ultrasonic signals, in another embodiment 530 and 532 send and receive infrared signals, and in another embodiment 530 and 532 send and receive radio signals. The receiver may also be absent.

A battery 540 provides power to the bookmarker 112. The bookmarker 112 is designed to fit conveniently into the human hand. In another embodiment, a hand-held personal computing device, such as the Palm™ hand-held device by 3Com®, may be configured to send and receive bookmarking signals from the audio and video notetaker 100.

Fig. 6 is a block diagram illustrating an embodiment of a mini docking station suitable for connecting the audio and video notetaker 100 to a personal computer. An audio and video notetaker mini docking station 600 contains an audio and video notetaker connector 610 for connecting the audio and video notetaker 100 to the mini docking station 600 for data transfer,

e.g. a Universal Serial Bus (USB) type, serial type or parallel type connector. The mini docking station 600 contains a power connection 620 suitable for charging the battery 390 of the audio and video notetaker 100. The mini docking station 600 also contains a personal computer connection 630 suitable for transferring data between an audio and video notetaker 100 and a personal computer, e.g. a standard USB connection. Data stored on the audio and video notetaker 100 is transferred to a personal computer for performing editing and playback of an audio and video meeting record. Alternately, the data may be edited "in place" on the device, when connected to a personal computer system using the audio and video notetaker editing software.

Fig. 7 is a flow diagram illustrating an embodiment of a method for processing and storing data in an audio and video notetaker system. As the data inputs 710, 720, and 730 are received, they are used in data analysis and data reduction. Both data analysis and data reduction are performed in real time on the incoming data inputs. The analyzed and reduced data is immediately stored 780. Data analysis, data reduction, and data storage 780 are performed closely proximate in time to when the data inputs 710, 720, and 730 were received. Thus when data input has completed, a final stored file 780 is ready to use, having already been analyzed and reduced.

The video input 710 and audio input 720 are received as a continuous stream of real-time data, for example, the audio and video notetaker system receives the data as it occurs during an ongoing meeting. The additional inputs 730 represent individual signals, such as individual bookmark signals sent repeatedly at discrete time intervals. In another embodiment, the inputs 730 also include whiteboard electronic input or a VGA-type data stream representing a presentation such as slides, charts or computer demos.

The data analysis process consists of an optional dewarping step 712, after which the video input 710, audio input 720, and other input 730 is fed into an analyzer 740. In one embodiment, dewarping 712 is performed on the raw video data 710. Captured images may be distorted by the wide-angle lens "fisheye" effect in a camera lens system that warps or otherwise distorts a received image before it is digitized. Dewarping corrects a distorted image via a corrective mapping procedure. The dewarping 712 process is calibrated individually for each camera used in the audio and video notetaker system. The process of dewarping video images is well known in the art. In another embodiment, the dewarping step 712 may be performed on the video data after selective video capture 750 has been performed. Dewarping is typically performed using dewarping software running on the audio and video

notetaker 100 CPU. However, optionally a dedicated dewarping hardware may be used to perform this function.

The video data 710 (optionally dewarped), the audio data 720, and the bookmarks and other input 730 are input into an analyzer 740. The analyzer 740 uses the inputs to analyze the received data and create a set of analyzer metadata containing information about the data streams 710, 720, and 730. The analyzer metadata represents a model of events occurring during the recording of the audio and video data, and this model of events contains information such as who spoke, when and where they spoke, and what was said during the meeting. The analyzer metadata is used for multiple purposes, including: (1) determining the region of interest for selective video capture and later viewing during playback, (2) audio selective capture decision making, (3) selecting video "mugshots" of the various meeting participants that will be linked to the audio recording of each participant's speech, (4) determining which of the particular meeting participants is speaking at any given time, and (5) for purposes of navigation and intelligent display of the recorded data. The analyzer 740 algorithms are robust with respect to a wide set of variables, including the number of participants, the lighting conditions in the room, the movement of the participants, the color of the background, and distracting events in the background.

The analyzer 740 may be implemented in a programmed digital signal processor (DSP) chip, or on a set of ASICs. Alternatively, the analyzer 740 may be implemented as a set of software modules to be run on the audio and video notetaker system CPU. Additional processing chips may be added as needed to the audio and video notetaker system to perform analyzer 740 processing. Processing may be shared between multiple modules.

The video input 710 and audio input 720 also feed directly into the data reduction process (which includes subsampling or other means of selective capture, as well as compression) after which they are stored. The video input 710 and audio input 720 are stored substantially close in time to when they were received by the audio and video notetaker system.

The bookmark signals and other input 730 are input into the analyzer 740, where they are incorporated into the analyzer metadata. Analyzer metadata derived from bookmarks and other inputs 730 is stored as auxiliary data that is associated with the audio and video data received at the same or substantially the same time as the input 730. In one embodiment, the auxiliary data is stored as a side file. The bookmark data is stored as a time offset reference,

indicating the time between when the audio and video notetaker system began recording and the time at which the bookmark signal was received. If the audio and video notetaker system bookmark signals include an indication of which participant transmitted the signal, the bookmark data also include a reference to the bookmark identifier, such as an identification number. In another embodiment, the bookmark data is stored as a frame reference number that refers to the compressed file portion that was recorded at the same time as the bookmark signal was received. In another embodiment, each bookmark data includes a time stamp referencing the portion of the audio and video data file recorded at the time the bookmark signal was received.

In another embodiment, the auxiliary data is stored in the same file as the audio and video data.

In one embodiment, a bookmark time delay parameter is set in the audio and video notetaker system. Bookmarks are associated with the data received at the time of bookmark receipt by the audio and video notetaker system minus the bookmark time delay parameter. The bookmark time delay parameter may be set to compensate for the human perceptual time delay that typically occurs between the time an event of interest occurs, and the time a participant actually sends a bookmark signal to the audio and video notetaker system. For example, if a particular person of interest begins speaking, it may take the person holding the bookmarker a few seconds before he or she hits the "bookmark" button to denote the event. The delay parameter may also be determined from other event analysis, such as pauses, the onset of speech, a change of speaker, etc. Bookmark delay parameters are included by the analyzer 740 in determining the time or frame number where the bookmark data is associated with the audio and video data file. In another embodiment, the bookmark delay parameter is not used during the initial storage of the bookmark data, but is used during editing and playback to adjust the bookmark data time or frame number association.

The analyzer 740 also produces additional metadata about the audio, video and other input data that is stored as auxiliary data, for example, as a linked side file to the stored audio and video data. The additional metadata is associated with the stored audio and video data by a time offset reference, an absolute time reference, or a frame number reference. However, analyzer metadata may or may not be generated substantially simultaneously with its associated audio and video data. The analyzer 740 "interprets" information about the events being recorded as they unfold. Certain analyzer metadata is determined and stored immediately. However, metadata developed later in time through an intelligent interpretation

process that corresponds to earlier-stored video and audio data is stored with the corresponding earlier data, and thus the storage 780 of metadata may not always occur linearly with time.

For example, the analyzer 740 is able to determine when a speaker change has occurred by identifying, for example, segments of silence in the audio recording and a change in the audio recording corresponding to a different voice pattern. The analyzer 740 may also determine a speaker change has occurred by identifying a changed location for the source of the audio, using microphone array triangulation. This analyzer metadata is recorded substantially simultaneously with the audio input 720.

However, it may take additional analysis for the audio and video notetaker system to be able to correlate individual speakers with particular sections of audio. The analyzer 740 uses, for example, voice recognition algorithms and patterns of movement in the video data to determine which speaker corresponds to which audio portions. As the event recording progresses, the analyzer 740 continues to receive additional information that may be used for speaker-to-audio correlation. Thus, the analyzer 740 may not be able to immediately correlate a particular speaker with a particular audio portion, but through an interpretation process will make the correlation later. Such "interpreted" analyzer metadata will be stored as auxiliary data correlated to the time of occurrence of the recorded event of interest, even though the information was interpreted at a later time.

High-resolution video and audio data requires a large amount of storage space. For example, the raw data received during a meeting may be approximately 1 gigabyte per hour of meeting time. The standalone version of the audio and video notetaker may contain approximately 8 gigabytes of memory, and a personal computer hookup might provide approximately 8 gigabytes more. An Ethernet connection would provide a great deal more storage, on the order of tens or even hundreds of gigabytes.

In order to conserve storage space and extend the recording capacity of the audio and video notetaker system, data reduction is performed on either one or both of the video 710 and audio 720 data before the data is stored. The data reduction process discards unnecessary or redundant data when possible. The data reduction process is structured to minimally impact the stored 780 audio and video resolution and quality. The data reduction process also includes the compression of the audio and video data.

The raw audio data 720 is input into an audio selective capture process 760, which also uses analyzer 740 metadata to perform selective capture of audio data. Audio data reduction is performed using audio selective capture to select and store important audio data while discarding background audio data not central to the events being recorded. In one embodiment, speech detection analysis is used with the raw audio input 720 to detect long pauses where no speech is occurring. Only audio portions containing speech data are stored in storage 780 and correctly associated with the corresponding video portions in time. In another embodiment, speaker detection analysis is used to identify particular speakers and avoid saving audio data associated with certain speakers. Additionally, in another embodiment a word spotting module detects and avoids saving all audio portions not close in time to the use of certain words or phrases. Key word spotting phrases may be preprogrammed in advance to ensure that only conversations regarding certain issues are saved. Each of these audio data reduction embodiments may be used alone or in combination with each other.

Dewarped video data 710 is input into a video selective capture process 750, which also uses analyzer 740 metadata to perform selective capture of video data. Fig. 8 is a block diagram illustrating an embodiment of a method for video data reduction. The method 750 shown in Fig. 8 uses selective video capture to select and save important video data while discarding background video data not central to the events being recorded.

In the embodiment shown in Fig. 8, a raw video data sample 710 is obtained by combining the views from four separate cameras at a particular point in time. For example, in a meeting there may be two participants A and C at a table. Camera view 810 includes only background with a briefcase; camera view 820 includes a participant A, a bookmarking device 112A and a coffee cup 126; camera view 830 includes only a background telephone; and camera view 840 includes a participant C and a bookmarking device 112C. Camera views 810 and 820 have an overlap region 812; camera views 820 and 830 have an overlap region 822, and camera views 830 and 840 have an overlap region 832. In one embodiment, each region 810, 820, 830 and 840 has a pixel resolution of 640 x 480 pixels. Each raw video data sample 710 therefore contains approximately 1.2 million pixels (480*640*4).

The data analysis process identifies regions of interest within the samples of the raw video data. Camera views 810 and 830 do not contain any regions of interest and no video data is saved from these views. A smaller region of interest 824 is selected from view 820. The region of interest 824 includes the image of participant A and bookmarker 112A, but eliminates the extraneous coffee cup image 126. A smaller region of interest 844 is selected

from view 840. The region of interest 844 includes participant C and bookmarker 112C, but eliminates extraneous background data. Only these regions of interest need be saved, and can thus be saved at higher resolution.

Analyzer metadata is used to determine what the regions of interest are in the raw video data. For example, as the analyzer 740 identifies faces and human figures in the video data 710, this metadata is used to determine regions of interest around participant faces. Additionally, the analyzer 740 identifies movement in the video data 710, and this metadata is similarly used to focus a region of interest around a video image with substantial movement.

In one embodiment, each region of interest 824 and 844 is saved as a portion of a tiled image 850. The tiled image 850 contains fewer pixels than raw video data sample 710, because it is composed of only the regions of interest. In one embodiment, the tiled image 850 has a pixel resolution of 720 x 480 pixels. Each tiled image frame 850 therefore contains a total of approximately 350,000 pixels. This represents a data reduction of approximately 70% over the 1.2 million pixels/frame of raw data sample 710.

In another embodiment, analyzer 740 metadata are used to identify event participants and save high-resolution data of the regions of interest surrounding individual figures. Background regions may be either discarded or saved at a reduced resolution. In embodiments using multiple cameras, duplicate images of overlapping regions may also be discarded. For example, region of interest 824 is partially located in the region of overlap 812. A region of interest chosen from camera view 810 would be chosen so as not to duplicate data saved from the camera overlap region 812. Subsampling and resampling techniques are well known in the art, and various other methods of data reduction will be evident to one of skill in the art.

Different embodiments of the selective video capture process 750 use different selection criteria in deciding what video data to save and what video data to discard. In one embodiment, subsampling or resampling methods are used to select a pattern of pixels to be discarded. For example, every other pixel from raw data sample 710 may be saved, thereby saving the entire field of view with a slightly reduced resolution. In another example, certain sets of pixels from raw data sample 710 may be blended together or averaged, and only the combined pixel values are saved.

After the video selective capture 750 and the audio selective capture 760, the resulting video and audio data streams are then compressed 770 to further reduce data storage requirements. In one embodiment, data compression 770 compresses the selected audio and

video data using a dedicated MPEG-2 compression chip. Video and audio compression schemes are well known in the art. It will be evident to one of skill in the art that various other data compression methodologies may be used to further compress the video and audio data after selective capture 750 and 760 have been performed.

It will be evident to one of skill in the art that the data reduction steps 750, 760 and 770 are performed to reduce data storage requirements for the audio and video notetaker system. In another embodiment, data reduction may be eliminated altogether. For example, if the audio and video notetaker 100 is connected via a port extender 110 to a large data storage system, data reduction may not be necessary if the connection bandwidth and external storage capacity is great enough. In this embodiment, the storage constraints of the audio and video notetaker system are removed, and the audio and video notetaker system no longer has to rely on its own internal storage mechanisms for storing the audio, video, bookmarks, and analyzer metadata data streams.

In one embodiment, data from the audio and video notetaker system is stored 780 in two separate, but linked, files. An MPEG file or similar file type is used to store the audio and video data. A separate, auxiliary metadata file is used to store bookmark data and other inputs 730. The separate auxiliary file also stores analyzer 740 metadata. The auxiliary file may be correlated to the MPEG file in several different ways, as described previously in regard to bookmark metadata. In one embodiment, the metadata is stored with an absolute or relative time stamp linking the metadata to a particular portion of audio and video data. For example, a speaker's identification is linked to the MPEG file by a time stamp corresponding to instances in time in which the speaker appears in the audio and video data. Although the speaker's identity may have been determined by the analyzer 740 somewhat later in time than the speaker appears in the video, the speaker identification metadata will be associated with the MPEG file at the point in time when the speaker appears in the MPEG file.

Fig. 9 is a block diagram illustrating an embodiment of the functionality of an analyzer 740 in an audio and video notetaker system. The analyzer 740 receives video 710, audio 720, and other inputs 730 and creates analyzer metadata 990 describing a model of the events being recorded through an intelligent interpretation process. The analyzer 740 is comprised of multiple different modules operating on the data received by the analyzer 740 and creating metadata 990.

Raw video input 710 is first prefiltered and preprocessed 902, in a manner allowing the analyzer 740 access to both the raw audio data 710 and the prefiltered and preprocessed video data. The video prefiltering and preprocessing module 902 performs computationally intensive processes on the video data, allowing different modules within the analyzer 740 to use the prefiltered and preprocessed video data without repeating computationally intensive processes among different modules. Similarly, raw audio input 720 is prefiltered and preprocessed 904 in a manner allowing the analyzer 740 access to both the raw audio data 720 and the prefiltered and preprocessed audio data.

For example, it is desirable for the analyzer 740 to be able to access both raw audio data and band pass filtered audio data. As the raw audio data 720 is input into the audio prefiltering and preprocessing module 904, a first buffer will store a segment of raw audio data. An additional second buffer will store the same segment of audio data after band pass filtering if requested by the analyzer 740. The analyzer 740 modules may access either of the two buffers for use in the analyzer 740 functions.

A set of data processing modules operate on the video and audio data. Each of these modules generates information that describes the events occurring during the meeting being recorded. The analyzer 740 combines the information generated by one module with information from additional modules that build upon the information already determined. The method shown in Fig. 9 for combining and verifying information is one embodiment of an analyzer 740. It will be evident to one of skill in the art that many other different methods of combining and verifying information may be used to create an analyzer 740. Additionally, it will be evident to one of skill in the art that the analyzer 740 may be designed to incorporate additional types of data processing modules as desired.

The functions of the various analyzer 740 modules will first be discussed separately. After the features of each individual module have been discussed, the interplay between the modules in one embodiment of the analyzer 740 will be discussed.

A visual tracker 910 uses video data to track people as they move around the environment being recorded by the audio and video notetaker system. Several different techniques for visual tracking are well known in the art, such as template-based tracking, edge-based tracking, and blob-based tracking. Template-based tracking is discussed in "Good Features to Track," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, by Jianbo Shi and Carlo Tomasi, pp. 593-600 (1994). Edge-based tracking is

discussed in "Tracking Non-Rigid Objects in Complex Scenes," Proceedings of the 4th International Conference on Computer Vision, by Daniel P. Huttenlocher, Jae J. Noh, and William J. Rucklidge, pp. 93-101 (1993). Blob-based tracking is discussed in "Incremental Focus of Attention for Robust Visual Tracking," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, by Kentaro Toyama and Gregory D. Hager, pp. 189-195 (1996).

A face detector 912 uses video data to detect human faces. Given a small window within a static image from the video stream, face detector 912 determines whether the window contains the pattern of a human face. Face detector systems trained on a large number of face and non-face examples can achieve good accuracy. Two techniques for face detection are given in "Human Face Detection in Visual Scenes," Advances in Neural Information Processing Systems Vol. 8, by Henry A. Rowley, Shumeet Baluja, and Takeo Kanade, pp. 875-881 (The MIT Press) (1996) and "A Statistical Approach to 3D Object Detection Applied to Faces and Cars," Henry Schneiderman, PhD Thesis CMV (2000).

A motion detector 914 uses video data to distinguish portions of the video stream that are static from those that are currently dynamic. Areas of motion within the video data indicate portions of the video in which meeting participants or other features of interest are likely to be located. A reference image frame is computed from previous video images, and the current image frame is subtracted from the reference frame to identify areas of dynamic change. A system using motion detection techniques is discussed in "Pfinder: Real-Time Tracking of the Human Body," IEEE Transactions on Pattern Analysis and Machine Intelligence Vol. 19 No. 7, by C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, pp. 780-785 (1997).

A skin color detector 916 uses video data to determine which pixels resemble human skin color. Although colors corresponding to human skin span a range of possible color values, and although other objects in the world map to the same colors as human skin, a skin color detector can effectively eliminate much of the image (in a typical environment) that does not look like human skin, while detecting most of the pixels corresponding to human skin. A technique for skin color detection is discussed in "Statistical Color Models with Application to Skin Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, by Michael J. Jones and James M. Rehg, pp. 274-280 (1999).

An acoustic localizer 918 uses audio data to determine the direction to the current speaker in terms of azimuthal and elevation angles. The acoustic localizer 918 uses a method involving: (1) matching signals from microphone pairs to estimate the time delay, (2) mapping the results from the various microphone pairs onto a common coordinate system, and (3) finding the direction from this combined result. Techniques for acoustic localization are discussed in "Use of the Crosspower-Spectrum Phase in Acoustic Event Location," IEEE Transactions On Speech and Audio Processing Vol. 5 No. 3, by Omologo and Svaizer (1997).

A human speech detector 920 uses audio data to distinguish segments of an audio signal containing human speech from segments not containing human speech. One technique for human speech detection that distinguishes speech from silence by relying on energy and number of zero crossings is discussed in "Digital Processing of Speech Signals," by L.R. Rabiner and R.W. Schafer (Upper Saddle River, New Jersey, Prentice Hall) (1978). Another technique distinguishing human speech from other sounds is discussed in "Detection of Human Speech using Hybrid Recognition Models," Proceedings of the IAPR International Conference on Pattern Recognition Vol. 2, by J.D. Hoyt and H. Wechsler, pp. 330-333 (1994). The human speech detector 920 operates on the audio data 720 and 904 to separate out audio portions containing speech. Audio portions containing speech are passed to the acoustic localizer 918 and a beamformer 930.

The beamformer 930 operates on audio data to enhance the signal-to-noise ratio of the speech of the current speaker. The beamformer 930 adds the signals of the various microphones of the audio and video notetaker system after first shifting them by non-constant amounts, based upon an estimate of the current speaker direction. Beamforming techniques are discussed in "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms," Journal of the Acoustical Society of America Vol. 78 No. 5, by Flanagan, Johnston, Zahn, and Elko (1985). The beamformer 930 improves the signal-to-noise ratio of the current speaker audio, and this enhanced audio signal is passed to a voice recognizer 960, a speech recognizer 970, and a word spotter 980.

The voice recognizer 960 uses audio data to determine which, if any, of the voices in a speaker database is the same speaker as the person speaking in a given segment of speech. The speaker database is automatically learned during the meeting, and may also use data from previous meetings if available. Acoustic localization and face tracking can be used to "bootstrap" the process of collecting speech samples, since these analysis modules provide independent indications of the identity of the current speaker. The voice recognizer 960 also

trains up the speaker database automatically, relying on the results of the analyzer 740 to segment the audio stream according to the various speakers. One technique for voice recognition performs dimensionality reduction by computing feature vectors from audio segments, and then matches the feature vectors against one another to determine if the segments of speech were uttered by the same speaker. This technique for voice recognition is discussed in "Voice Recognition," by Richard L. Klevans and Robert D. Rodman (Boston: Artech House) (1997). Each recorded segment of speech is associated with a particular speaker.

The speech recognizer 970 uses audio data to provide a transcription of the human speech in the audio data. Two different commercially available software packages for speech recognition are Dragon Systems® Inc.'s NaturallySpeaking® and IBM®'s ViaVoice™.

The word spotter 980 uses audio data to detect when a word in a precomputed dictionary is spoken. Word spotting is similar to speech recognition, except that the dictionary contains only high-level keywords and excludes common words such as articles and prepositions. In one embodiment, the speech recognizer 970 and the word spotter 980 are allowed to run off-line, for example during recording playback and editing, because the transcription and keywords that these modules provide are not used by any other modules within the analyzer 740.

A person detector 940 combines the results of the face detector 912, the motion detector 914, the skin color detector 916 and the acoustic localizer 918 to determine which pixel regions of the recorded environment include people. The person detector 940 determines the number of participants in the recorded meeting, their locations, and their identifications (i.e., associates a visually identified participant as a particular audio speaker and also tracks the identity of a participant across frames). As computational resources are available, the person detector 940 calls the face detector 912 to verify the person detector 940 results.

The participants detected by the person detector 940 are passed to a person maintainer 950, which tracks the participants as they move about the environment using the visual tracker 910. Periodically, as computational resources are available, the person maintainer 950 calls the face detector 912 to compare the voice and face pattern of the current faces with those of previous faces to verify the person maintainer 950 tracking results.

The results of the person maintainer 950, the voice recognizer 960, the speech recognizer 970 and the word spotter 980 are used to create a set of metadata 990. As shown in

Fig. 7, the metadata is used in video selective capture 750, audio selective capture 760, and also stored 780 as a metadata file. For example, the metadata 990 contains information such as the number of participants present during the events being recorded, each participant's location at different times in the video stream, mugshots of the faces of each speaker, the identity of each speaker correlated to the speaker's audio, metadata about the selective capture of video and audio, and the time stamp, time offset, or MPEG frame associated with each bookmark. The metadata 990 is continuously updated and verified with new information gathered as events progress and additional raw data is available for processing and analysis.

Additional modules may be added to the analyzer 740 to operate on the bookmarks and other user inputs 730 as well as other aspects of the audio and video data. In the embodiment shown in Fig. 9, a slide transition detector 922 operates on VGA slide presentation input to detect when a new slide is being viewed. The slide transition detector 922 triggers the audio and video notetaker system to record each new slide, without storing repetitive images of the same slide.

It should be understood that the embodiments described herein can be implemented using any appropriate combination of hardware and/or software and that the embodiments described herein are provided for the sake of example and are not to be taken in a limiting sense. It should also be understood that if the embodiments described here are implemented in software, the software programs are stored in a memory, or on a computer readable medium, and are executed by a processor or processors, as appropriate.

Although the invention has been described in considerable detail with reference to certain embodiments, other embodiments are possible. As will be understood by those of skill in the art, the invention may be embodied in other specific forms without departing from the essential characteristics thereof. For example, video recording may be performed using one of several different camera embodiments. Additional metadata interpretation and analysis modules may be implemented within the analyzer to perform additional processing on the audio, video and bookmarking data inputs. Accordingly, the present invention is intended to embrace all such alternatives, modifications and variations as fall within the spirit and scope of the appended claims and equivalents.

We claim:

1. A system for the capture and analysis of audio and video data, comprising:

a multidirectional audio recorder for recording audio data;

a multidirectional video recorder for recording video data in synchrony with the audio data;

an event indicator for sending a bookmark signal at about an event time during the recording of the audio and video data; and

a bookmark module communicatively coupled to the event indicator to receive the bookmark signal, and associate a bookmark data with a portion of at least one of audio data recorded at about the time of the event, or video data recorded at about the time of the event.

2. The system of claim 1 wherein the system further includes:

a digital memory coupled to receive the recorded audio and video data, and store the audio and video data and the bookmark data in a format adapted for random access to the audio and video data associated with the bookmark data.

3. The system of claim 2 wherein the digital memory is included in a computer communicatively coupled to the system.

4. The system of claim 2 wherein the bookmark data are stored in an auxiliary file including a time stamp associated with the video data.

5. The system of claim 2 wherein the bookmark data are stored in an auxiliary file including a frame reference associated with the video data.

6. The system of claim 1, further including:

an analyzer adapted to perform real-time audio analysis using the audio data to produce metadata representing a model of events occurring during the recording of the audio and video data.

7. The system of claim 6 wherein the metadata are stored in an auxiliary file including a time stamp associated with the video data.

8. The system of claim 6 wherein the metadata are stored in an auxiliary file including a frame reference associated with the video data.

9. The system of claim 6 wherein the analyzer further comprises:

an acoustic localizer to determine the location of origin of a sound in the recorded audio data.

10. The system of claim 6 wherein the analyzer further comprises:

a human speech detector for distinguishing human speech from other audio data in the recorded audio data.

11. The system of claim 6 wherein the analyzer further comprises:

a speech recognition module for mapping recorded audio data of human speech to text.

12. The system of claim 6 wherein the analyzer further comprises:

a voice recognizer for identifying a speaker with a particular portion of the recorded audio data.

13. The system of claim 6 wherein the analyzer further comprises:

a word spotting module for detecting individual words in the recorded audio data.

14. The system of claim 6 wherein the metadata includes:

speaker identification metadata identifying a particular speaker with a particular portion of audio data;

audio selective capture metadata identifying selected portions of the audio data that will be stored; and

audio event detection metadata associating a particular portion of audio data with a particular event.

15. The system of claim 1, further including:

an analyzer adapted to perform real-time video analysis using the video data to produce metadata representing a model of events occurring during the recording of the audio and video data.

16. The system of claim 15 wherein the metadata are stored in an auxiliary file including a time stamp associated with the video data.

17. The system of claim 15 wherein the metadata are stored in an auxiliary file including a frame reference associated with the video data.

18. The system of claim 15 wherein the analyzer further comprises:

a skin color detector for identifying human skin color pixels within the video data.

19. The system of claim 15 wherein the analyzer further comprises:

a motion detector for identifying motion within the video data.

20. The system of claim 15 wherein the analyzer further comprises:

a face detector for detecting human faces within the video data.

21. The system of claim 15 wherein the analyzer further comprises:

a visual tracker for tracking the motion of human figures as they move about within the video data.

22. The system of claim 15 wherein the analyzer further comprises:

a person detector for correlating video data and audio data to determine where human figures are located within the video data.

23. The system of claim 15 wherein the analyzer further comprises:

a person maintainer for correlating video data and audio data over time to track human figures determined to be located within the video data.

24. The system of claim 15 wherein the metadata includes:

participant identification metadata identifying a participant with a particular portion of video data;

video selective capture metadata identifying selected portions of the video data that will be stored; and

video event detection metadata correlating a particular portion of video data with a particular event.

25. The system of claim 1, further including:

an analyzer adapted to perform real-time video analysis using the video data to produce metadata representing a model of events occurring during the recording of the audio and video data; and

the analyzer further adapted to perform real-time audio analysis using the audio data to produce metadata representing a model of events occurring during the recording of the audio and video data.

26. The system of claim 1 wherein the audio recorder is a multidirectional audio recorder.

27. The system of claim 26 wherein the multidirectional audio recorder comprises a plurality of microphones.

28. The system of claim 1 wherein the video recorder includes a multidirectional camera system that collectively covers a substantially 360° field of view.

29. The system of claim 28 wherein the multidirectional camera system is a plurality of cameras arranged in a substantially ring-shaped arrangement.

30. The system of claim 28 wherein the multidirectional camera system is an omnidirectional video camera.

31. The system of claim 30 wherein the omnidirectional video camera includes an omnidirectional camera using a curved mirror.

32. The system of claim 30 wherein the omnidirectional video camera includes a plurality of cameras and a pyramid-shaped mirror, each camera facing one side of the pyramid-shaped mirror.

33. The system of claim 1 wherein the video recorder includes a multidirectional camera system wherein a first camera is pointed in a first direction and a second camera is pointed in a second direction substantially 180° from the first direction.

34. The system of claim 1 wherein the video recorder includes a wide-angle camera system.

35. The system of claim 1 wherein the bookmark signal is an infrared signal.

36. The system of claim 1 wherein the bookmark signal is a radio signal.

37. The system of claim 1 wherein the bookmark signal is an ultrasonic signal.

38. The system of claim 1 wherein the bookmark signal is a signal sent via a wire.

39. The system of claim 1 wherein the bookmark signal is sent when a system user performs a manual action on the event indicator.

40. A portable device for the dynamic capture, analysis and storage of audio and video data, comprising:

a plurality of microphones arranged to provide multidirectional audio recording of audio data including one or more speakers;

one or more cameras to provide a multidirectional field of view video data recording;

one or more ports for receiving bookmark signals, which signals indicate the occurrence of an event and are associated with a portion of at least one of the audio or video data recorded at the time each bookmark signal was received;

one or more processing units for receiving the audio data and performing real-time analysis on the audio data to identify distinct speakers in the audio data, and further for receiving the video data and performing real-time analysis on the video data to identify portions of the video data associated with distinct speakers; and

a digital memory coupled to the audio processing unit and the video processing unit for storing the audio data, the video data, the bookmarks and the analysis data.

41. The portable device of claim 40, wherein the portable device is dimensioned to fit into a pie plate.

42. The portable device of claim 40, wherein the portable device includes an accessible control button for pausing and resuming recording of audio and video data by the portable device.

43. The portable device of claim 40, wherein the portable device includes a display indicating the recording status of the portable device.

44. A system for the dynamic capture and analysis of audio and video data, comprising:

a recorder adapted to multidirectionally record audio data and video data and receive bookmark signals which indicate the occurrence of an event and are associated with a portion of at least one of the audio or video data recorded at about the time each bookmark signal was received;

an port extender connected to the recorder and adapted to receive additional external signals; and

one or more portable bookmarkers adapted to create a bookmark signal.

45. The system of claim 44, wherein the port extender allows the external signals to pass through the port extender, and further transmits the external signals to the recorder.

46. The system of claim 44, wherein the portable bookmarkers are dimensioned to fit into the palm of a human hand.

47. The system of claim 44, wherein the portable bookmarkers are hand-held computing devices.

48. The system of claim 44, wherein the external signal is a video graphics array signal.

49. The system of claim 44, wherein the external signal is a universal serial bus signal.

50. The system of claim 44, wherein the external signal is a National Television Standards Committee signal.

51. The system of claim 44, wherein the external signal is transmitted using a wireless configuration.

52. In a system including digital, multidirectional audio and video recorders, a method for the dynamic capture and analysis of audio and video data, comprising:

receiving video data providing a multidirectional field of view;

receiving multidirectional audio data correlated in time with the video data;

analyzing the audio and video data in real time to create metadata correlating distinct speakers in the audio data with distinct images of speakers in the video data and tracking the speakers over time;

selecting portions of the audio data for recording;

selecting portions of the video data for recording;

recording the selected portions of the audio and video data; and

recording the metadata.

53. The method of claim 52, further including receiving a bookmark signal indicating a contemporaneous event of interest and indexing a corresponding bookmark data with the audio data and the video data.

54. The method of claim 53, further including storing the video data, the audio data, the bookmark data and the metadata in a digital memory in a format adapted for random access.

55. The method of claim 52, wherein the step of analyzing further comprises determining the number of people represented in the audio data.

56. The method of claim 52, wherein the step of analyzing further comprises determining the number of people represented in the video data.

57. The method of claim 52, wherein the step of analyzing further comprises locating faces in the video data.

58. The method of claim 52, wherein the step of analyzing further comprises dewarping the video data.

59. The method of claim 52, wherein the step of selecting portions of the audio data for recording further comprises:

analyzing the audio data and discarding audio portions that do not contain human speech; and

compressing the non-discarded audio data.

60. The method of claim 52, wherein the step of selecting portions of the video data for recording further comprises:

subsampling the video data to reduce image resolution, and

compressing the subsampled video data.

61. The method of claim 52, wherein the step of selecting portions of the video data for recording further comprises:

using the metadata to identify regions of interest in the video data; and

compressing the identified regions of interest in the video data.

62. In an audio and video notetaker system including multidirectional audio and video recording, a method for using an audio and video notetaker, the method comprising:

turning on the audio and video notetaker in a location in which meeting participants are visible and audible;

holding a meeting; and

generating bookmarks associated with events in the meeting that are stored as data correlated to the audio and video data recorded.

63. The method of claim 62 further comprising:

coupling the audio and video notetaker to an electronic whiteboard; and

capturing signals from the electronic whiteboard.

64. The method of claim 62 further comprising:

coupling the audio and video notetaker to a remote video graphics array signal; and

capturing the video graphics array signal.

65. The method of claim 62 further comprising:

coupling the audio and video notetaker to a remote video feed; and

capturing signals from the remote video feed.

66. The method of claim 62 further comprising:

coupling the audio and video notetaker to a projector display; and

capturing signals from the projector display.

67. The method of claim 62 further comprising:

pausing the operation of the audio and video notetaker during a portion of the meeting.

Figure 1A

Figure 1B

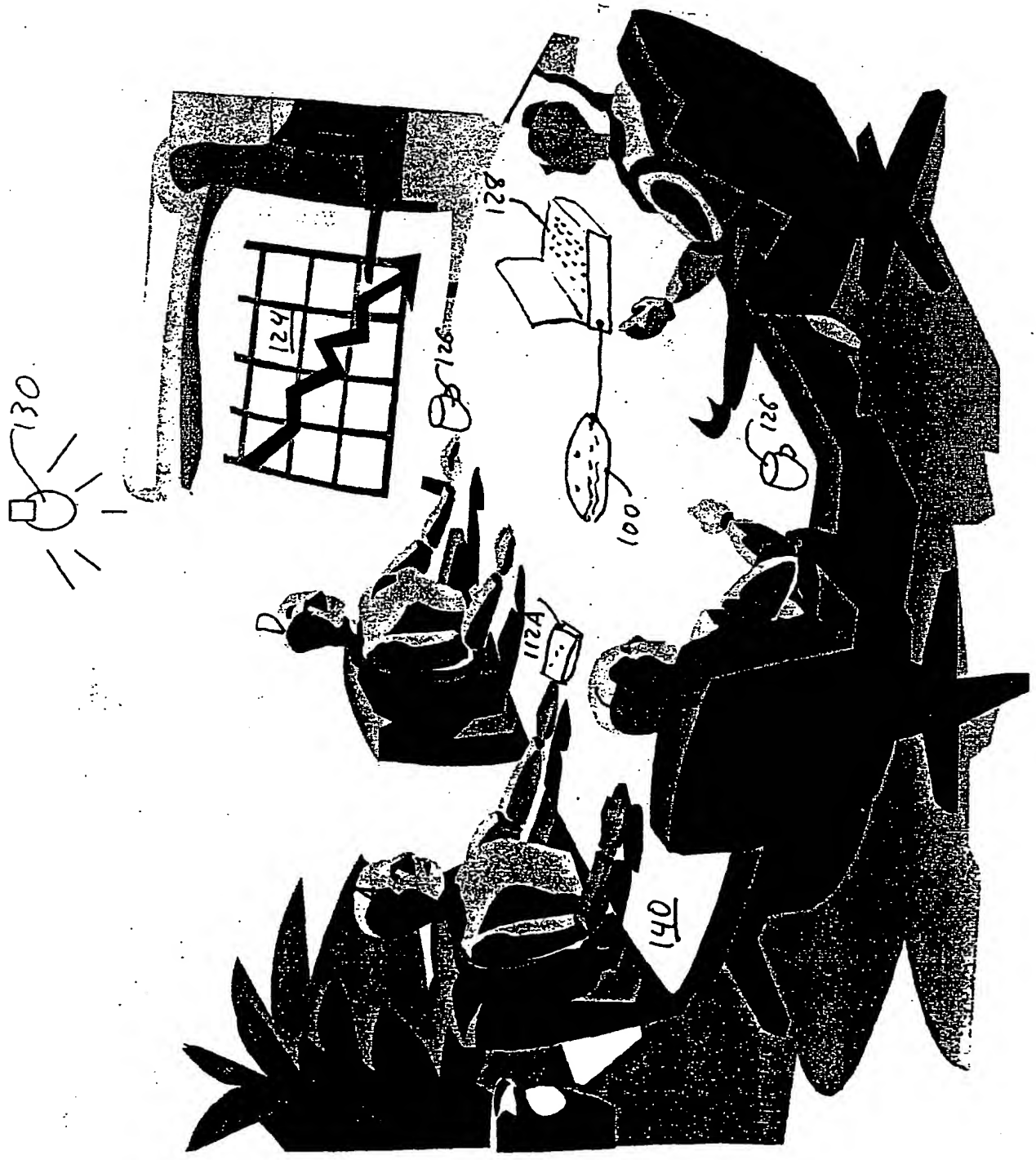

Figure 1C

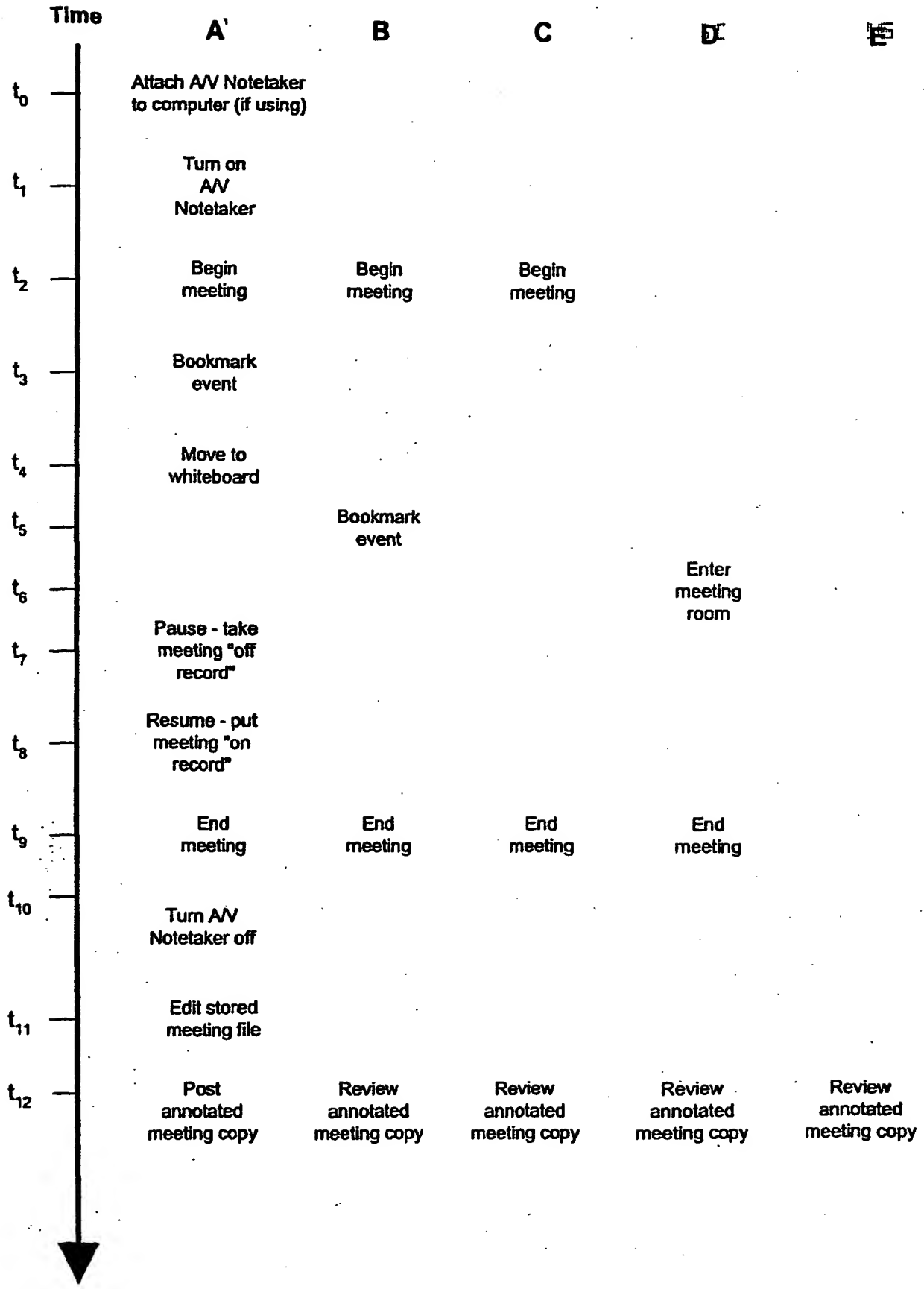
Time
A
B
C
D
E
t_0
Attach A/V Notetaker to computer (if using)
t_1
Turn on A/V Notetaker
t_2
Begin meeting
Begin meeting
Begin meeting
t_3
Bookmark event
t_4
Move to whiteboard
t_5
Bookmark event
t_6
Enter meeting room
t_7
Pause - take meeting "off record"
t_8
Resume - put meeting "on record"
t_9
End meeting
End meeting
End meeting
End meeting
t_{10}
Turn A/V Notetaker off
t_{11}
Edit stored meeting file
t_{12}
Post annotated meeting copy
Review annotated meeting copy
Review annotated meeting copy
Review annotated meeting copy
Review annotated meeting copy

Figure 1D

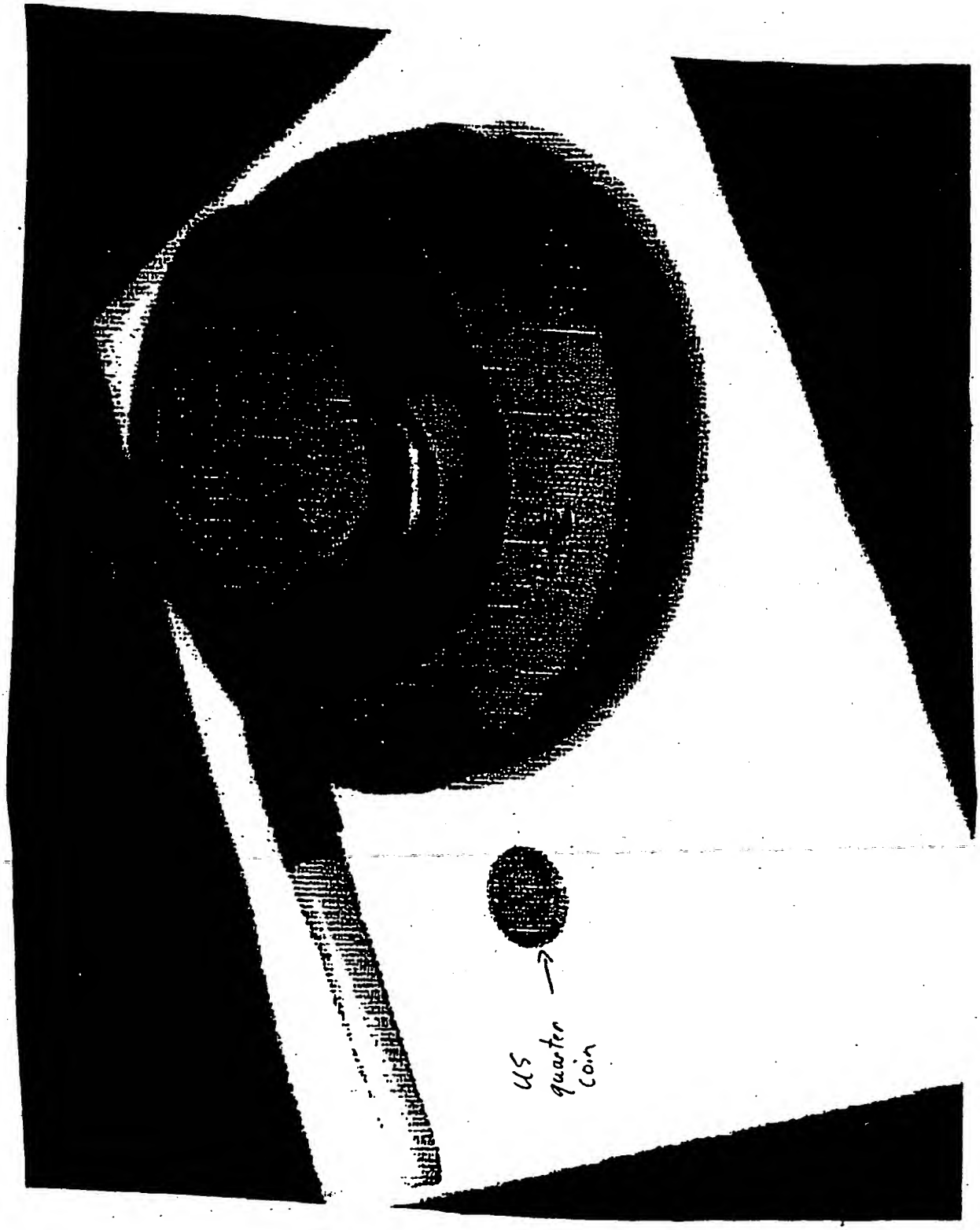

Figure 2A

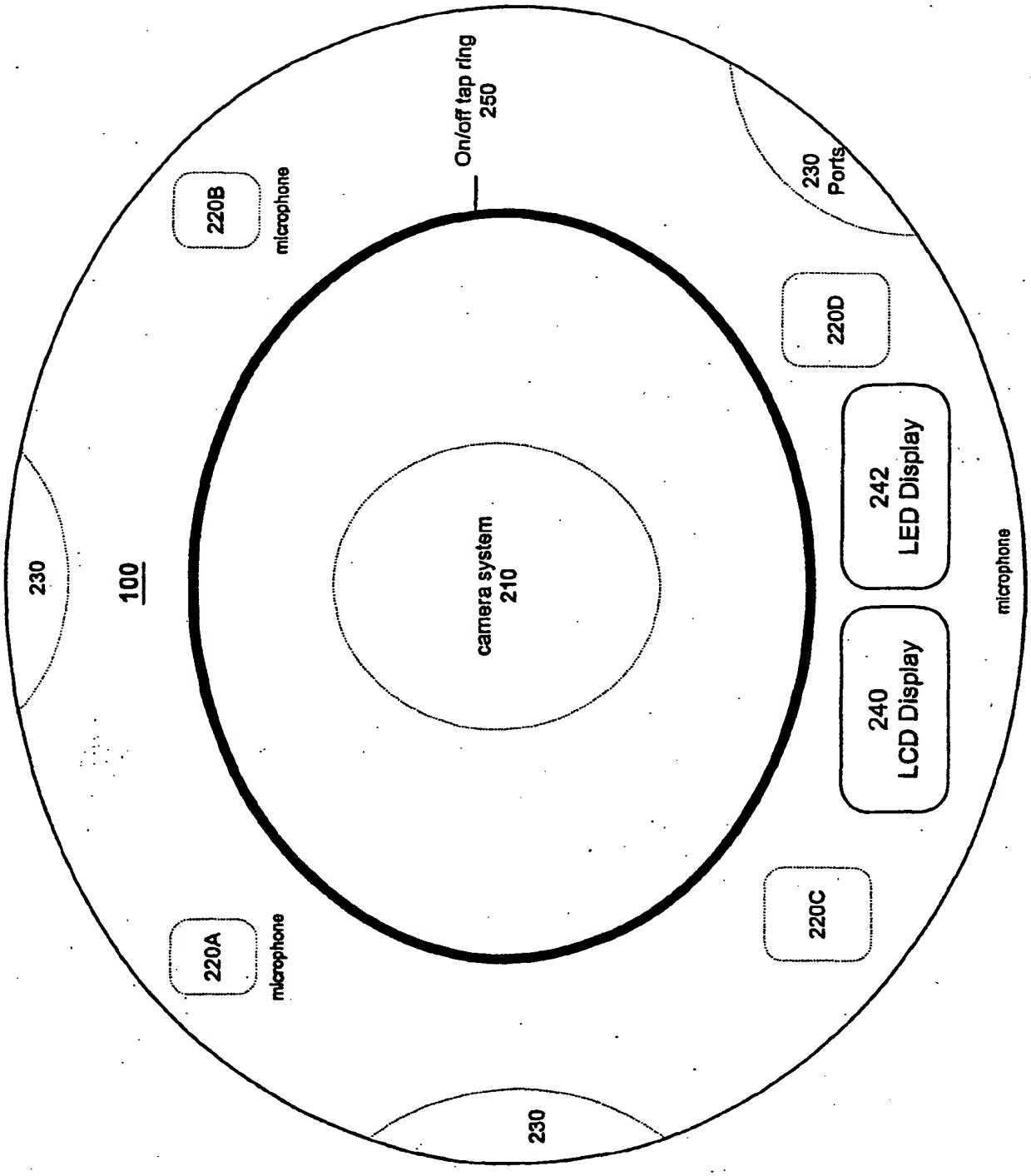
On/off tap ring
250
230
Ports
220B
microphone
220D
242
LED Display
240
LCD Display
camera system
210
100
230
220A
microphone
220C
230
microphone

Figure 2B

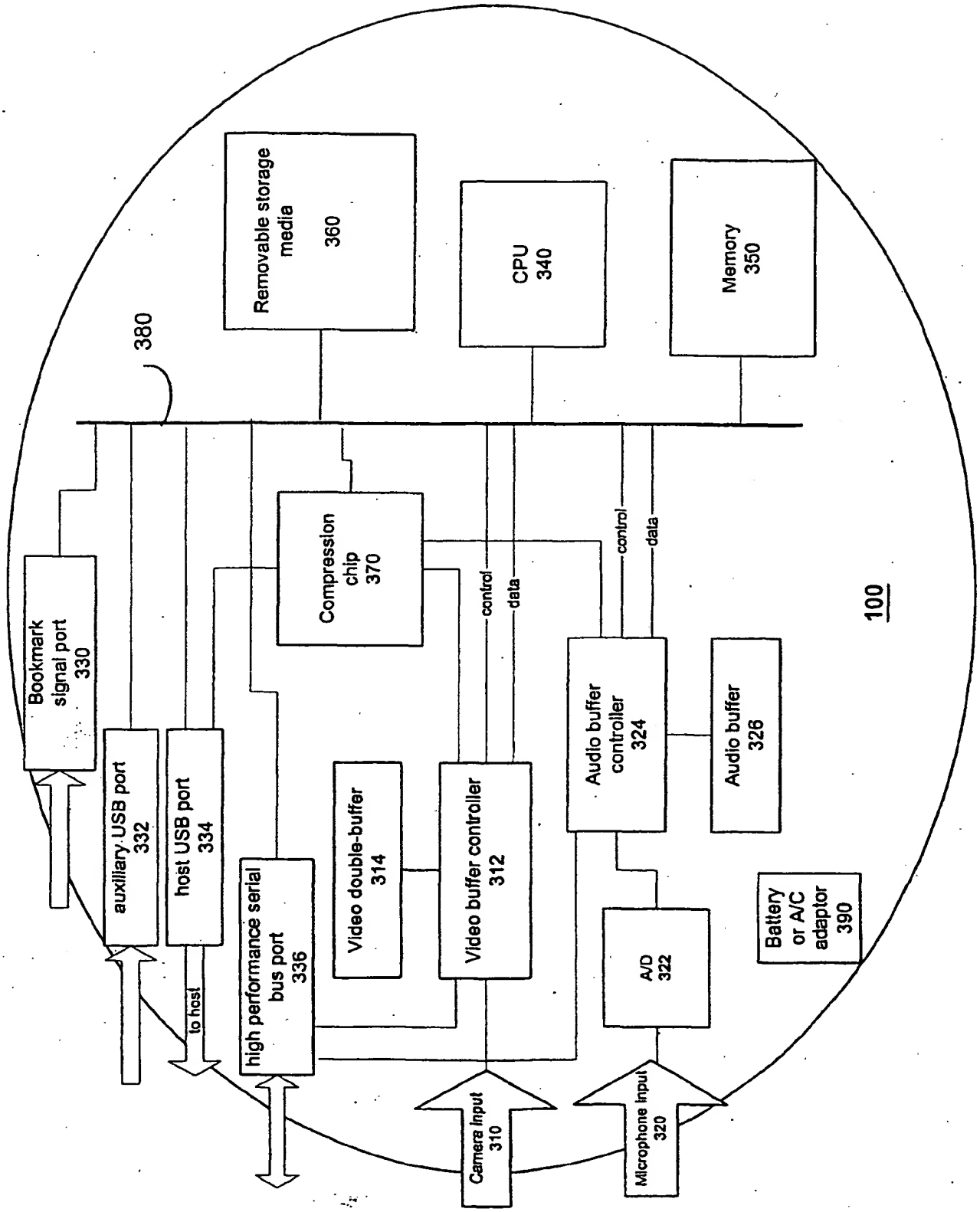
Removable storage media 360
CPU 340
Memory 350
380
Bookmark signal port 330
auxiliary USB port 332
host USB port 334
high performance serial bus port 336
to host
Compression chip 370
control
data
Video double-buffer 314
Video buffer controller 312
Audio buffer controller 324
Audio buffer 326
A/D 322
Battery or A/C adaptor 390
Camera input 310
Microphone input 320
100

Figure 3

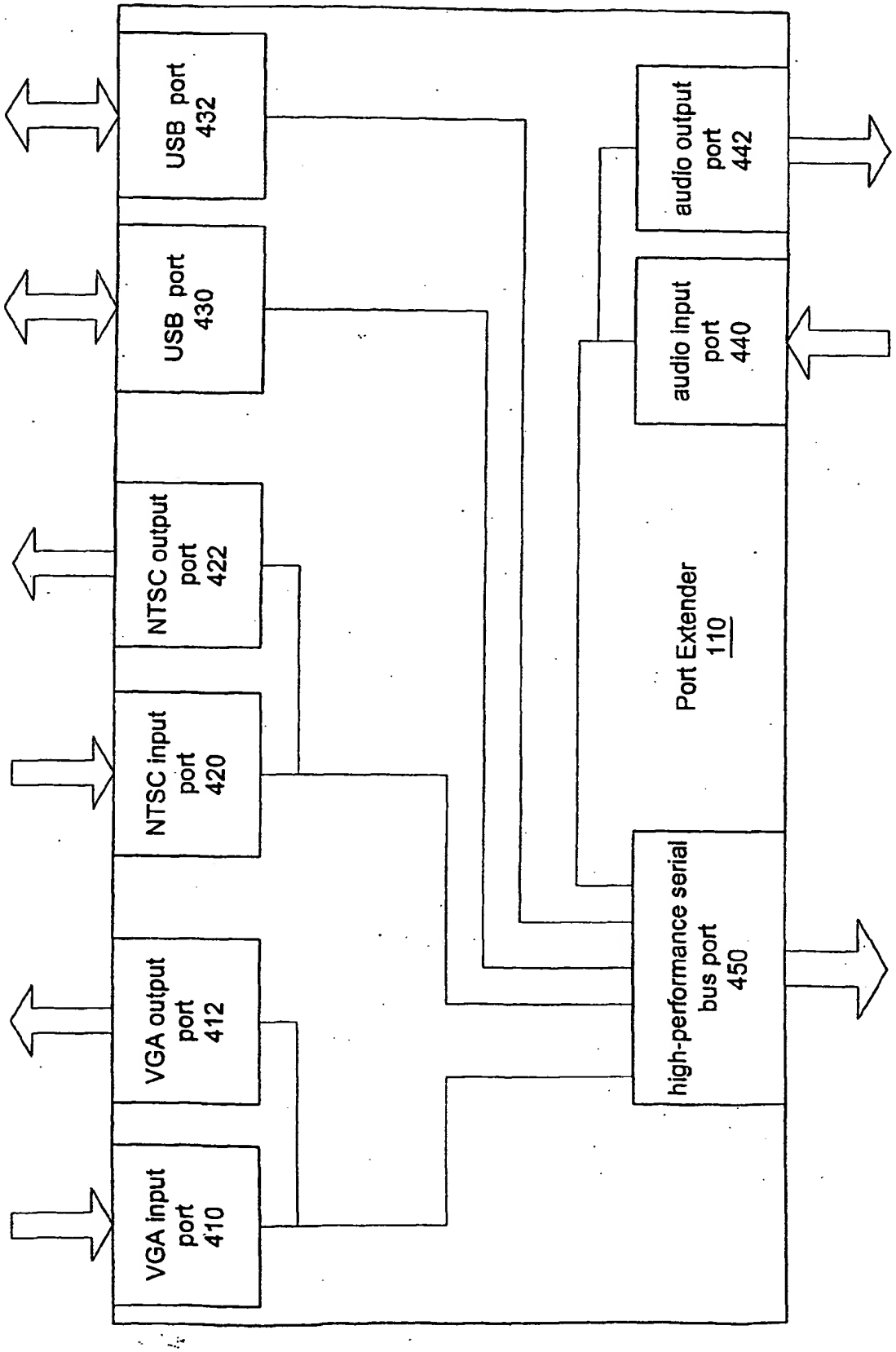
USB port 432
USB port 430
NTSC output port 422
NTSC input port 420
VGA output port 412
VGA input port 410
audio output port 442
audio input port 440
high-performance serial bus port 450
Port Extender 110

Figure 4

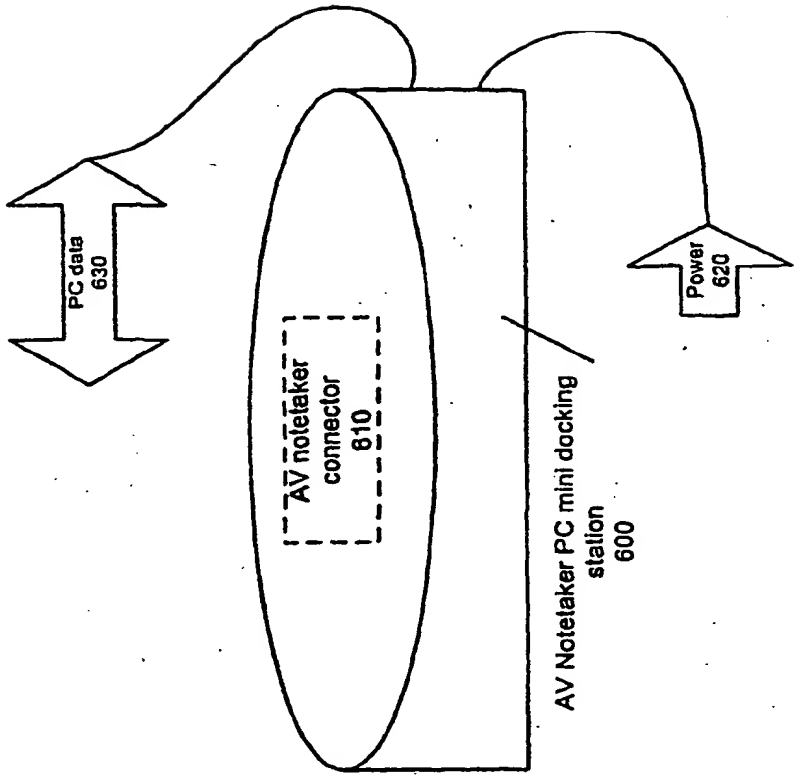

Figure 6

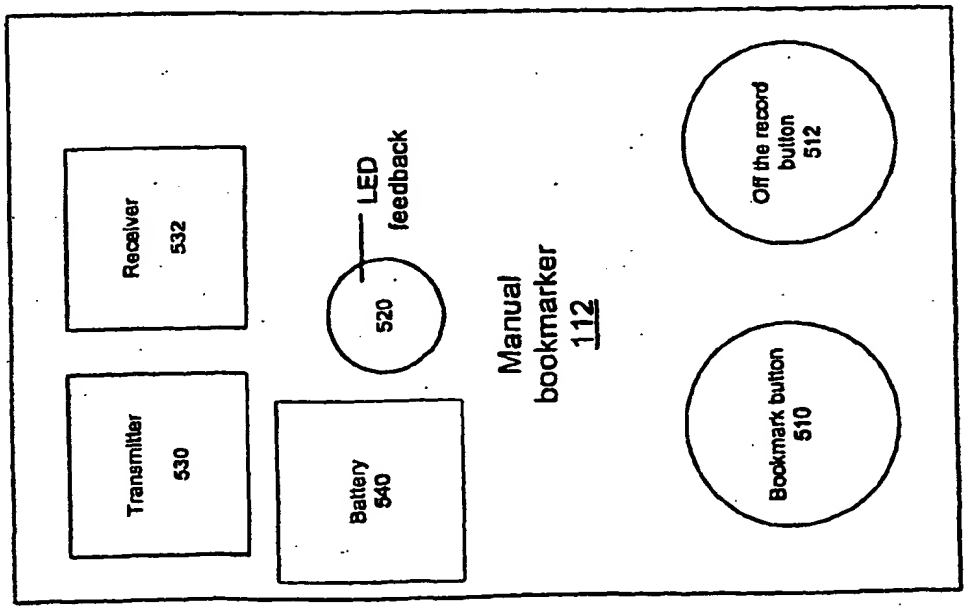

Figure 5

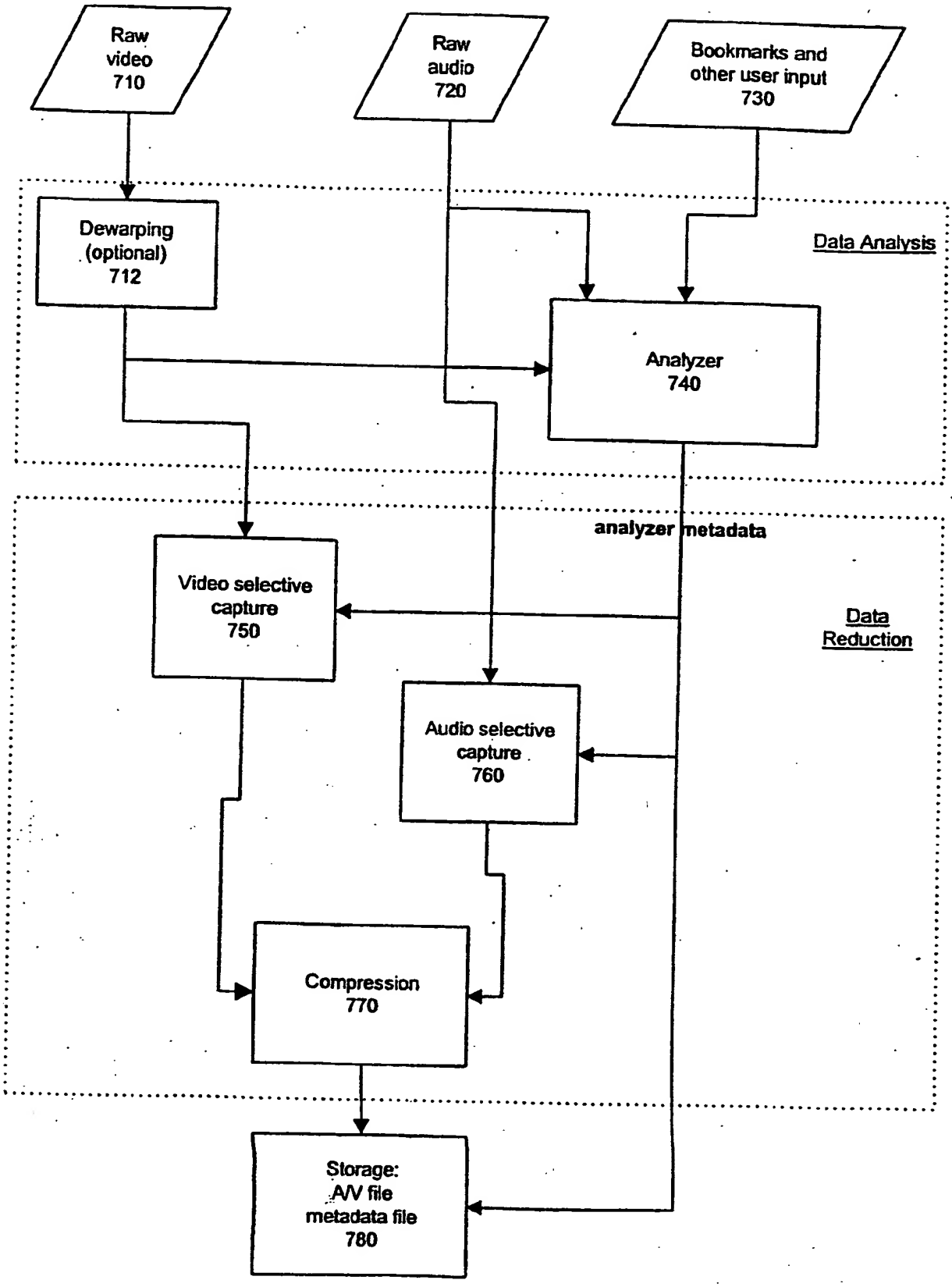
Raw video 710
Raw audio 720
Bookmarks and other user input 730
Dewarping (optional) 712
Data Analysis
Analyzer 740
analyzer metadata
Video selective capture 750
Data Reduction
Audio selective capture 760
Compression 770
Storage: A/V file metadata file 780

Figure 7

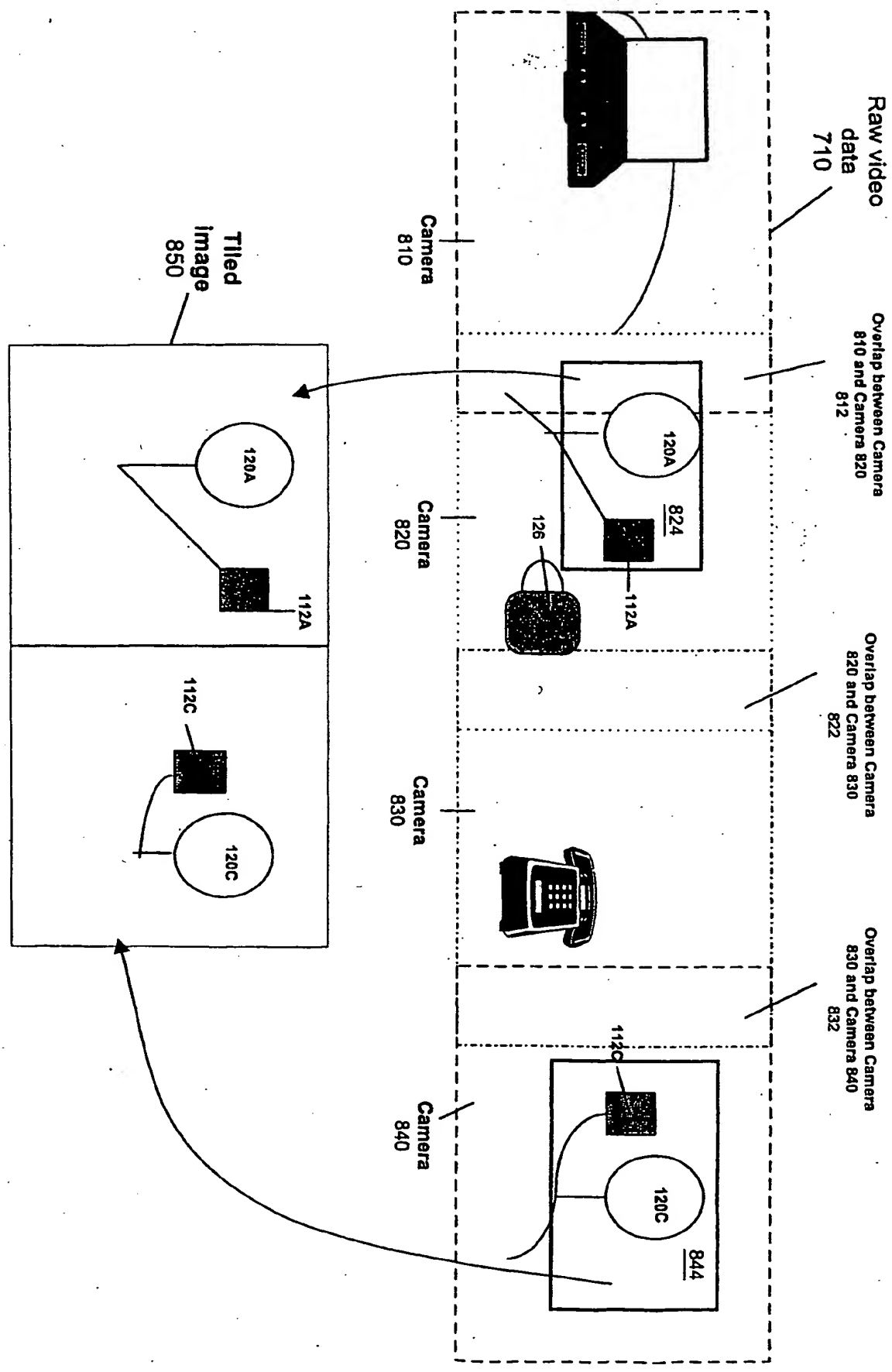
Raw video data 710
Camera 810
Overlap between Camera 810 and Camera 820 812
Camera 820
Overlap between Camera 820 and Camera 830 822
Camera 830
Overlap between Camera 830 and Camera 840 832
Camera 840
Tiled Image 850
120A
824
126
112A
112C
120C
844
750

Figure 8

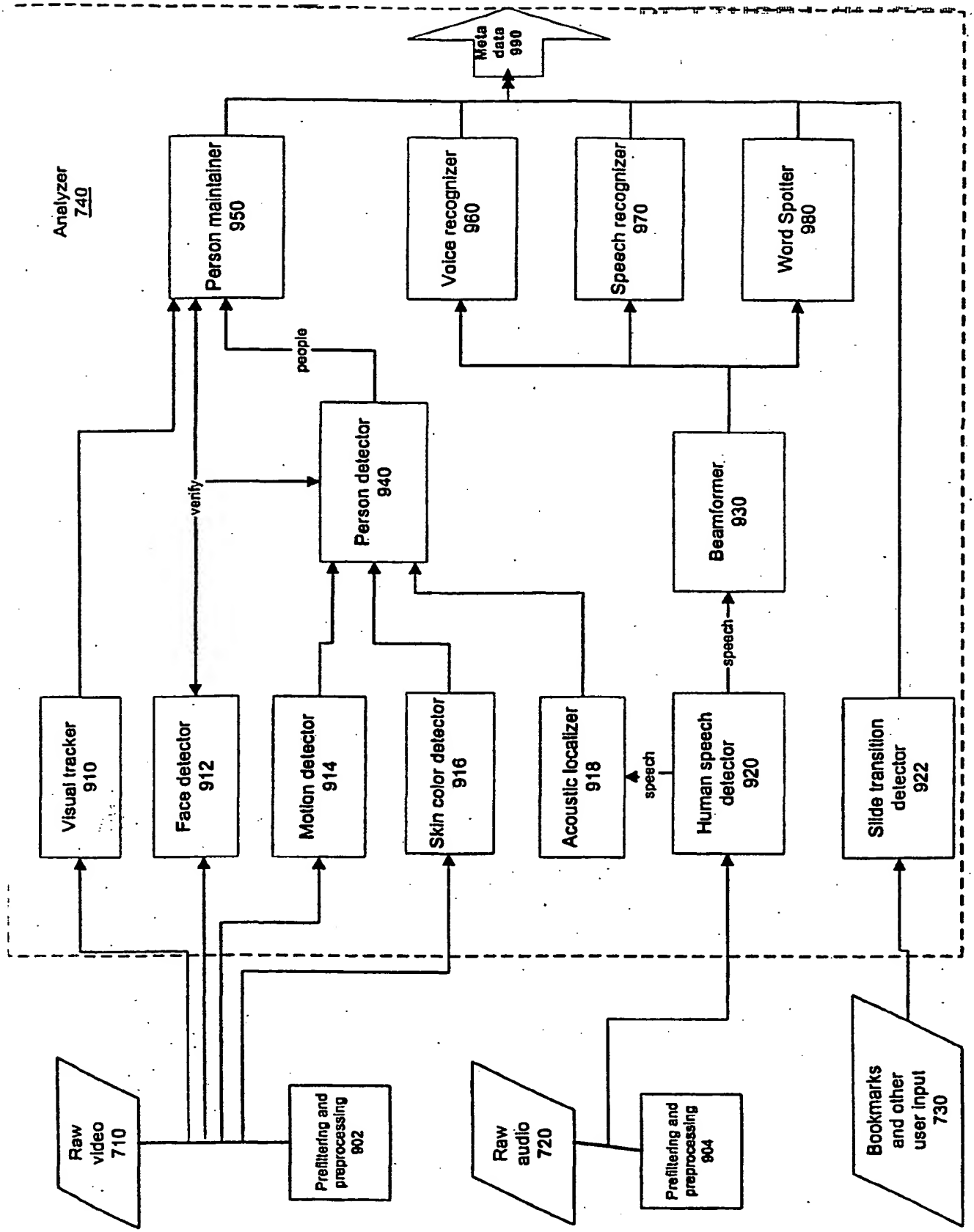
Meta data 990
Analyzer 740
Person maintainer 950
Voice recognizer 960
Speech recognizer 970
Word Spotter 980
people
verify
Person detector 940
Beamformer 930
speech
Visual tracker 910
Face detector 912
Motion detector 914
Skin color detector 916
Acoustic localizer 918
speech
Human speech detector 920
Slide transition detector 922
Raw video 710
Prefiltering and preprocessing 902
Raw audio 720
Prefiltering and preprocessing 904
Bookmarks and other user input 730

Figure 9